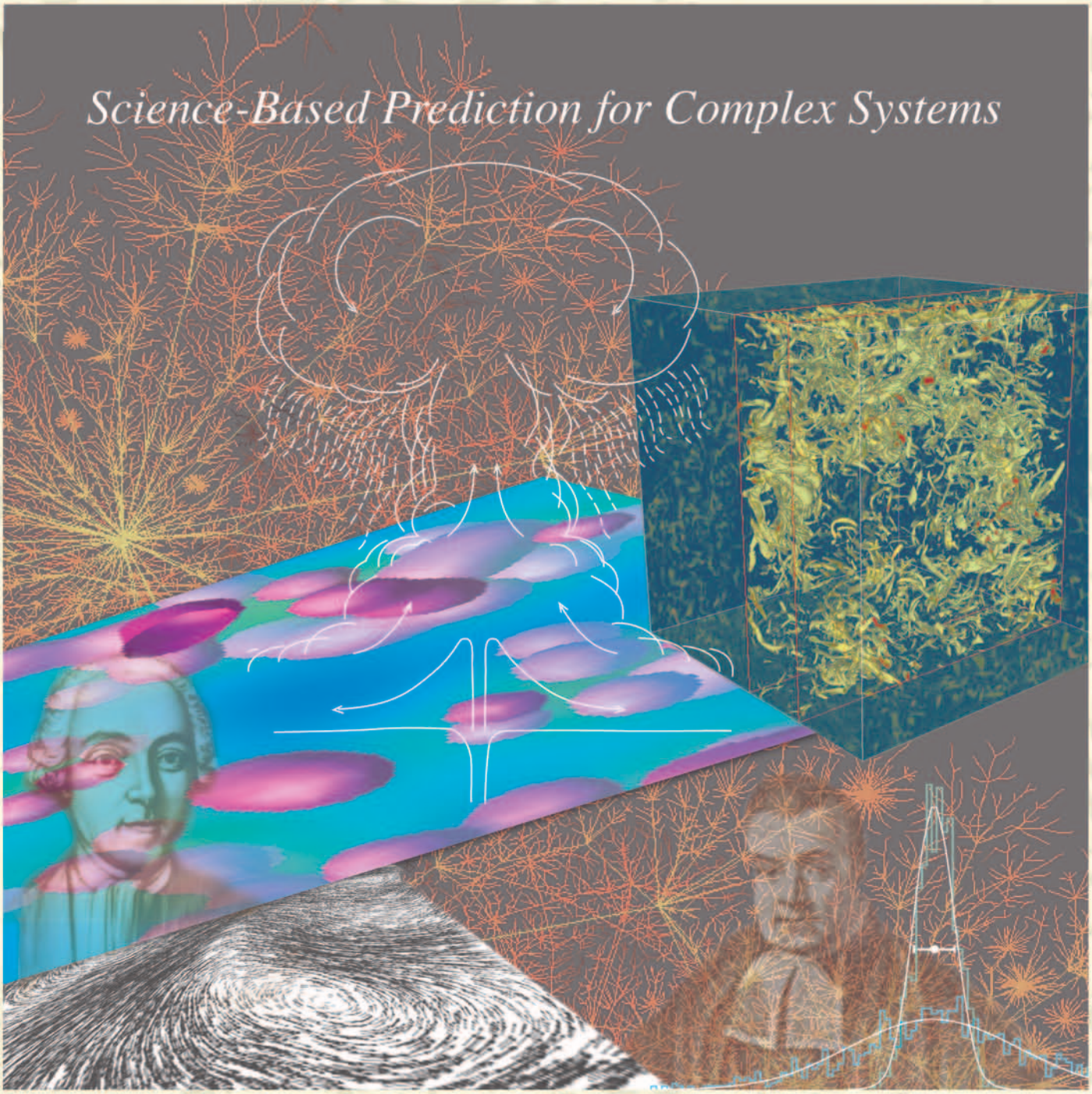
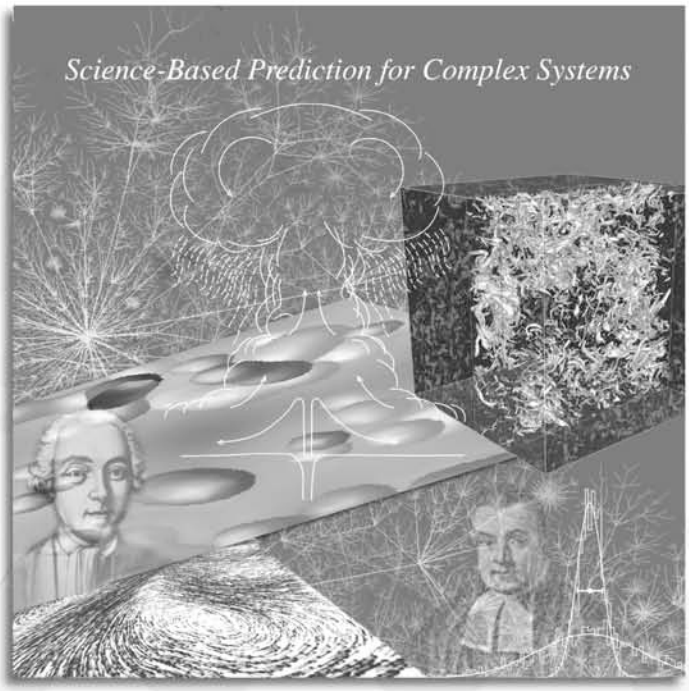
Los Alamos Science

LOS ALAMOS NATIONAL LABORATORY



Number 29 2005

Los Alamos NATIONAL LABORATORY EST. 1943



On the Cover

Complex systems come in many forms. Those on the cover were imaged through observation and computer simulation. The scale-free network filling the background shows the connections on the Internet at an instant in time. The partially ionized atoms (pink spheres) in a blue sea of free electrons (left and center) represent a quantum molecular dynamics simulation of the "warm" dense matter found in giant planets. The periodic box containing the cascading swirls of decaying turbulence (middle right) shows results from one of the largest simulations ever completed on the Los Alamos Advanced Simulation and Computing Q supercomputer. A single black-and-white swirl (bottom left) in turbulent flow shows velocity data acquired with high-power pulsed lasers and computer-automated data acquisition systems. Finally, an artist's drawing suggests the power and shape of a huge volcanic eruption (center), not unlike those of a nuclear explosion.

Attempts to predict the behaviors of such diverse systems rest not just on the power of modern supercomputers, but also on the inventiveness of the human mind and the edifice of mathematical and physical principles developed over centuries. Representative for this volume is the prolific mathematician Leonhard Euler (1707–1783), pictured at lower left. Euler wrote down the first fluid equations of motion and invented the field of graph (or network) theory. Across from Euler is Reverend Thomas Bayes (1702–1761), who was the first to use probability for inductive reasoning. Bayes' theorem for conditional probabilities (actually written down in present-day form by P. S. de Laplace) lays out the fundamental rule of statistical inference for determining the most likely behavior of complex, many-component systems. Bayesian analysis was used to reach a dramatic reduction in uncertainty for predicting nuclear fission-related processes, as illustrated by the new and old probability curves (sharply peaked and broad, respectively, in the lower right corner) for the nuclear criticality of Jezebel, a Los Alamos nuclear assembly for integral experiments.

Editor
Necia Grant Cooper

Managing Editor
Ileana G. Buican

Designer
Gloria E. Sharp

Illustrators
Andrea J. Kron
Christopher D. Brigman

Editorial Support
Brian H. Fishbine

Composition and Distribution
Joy E. Baker

Composition Support
Joyce A. Martinez

Photographers
Richard C. Robinson

Printing Coordination
Guadalupe D. Archuleta

Address mail to
Los Alamos Science
Mail Stop M711
Los Alamos National Laboratory
Los Alamos, NM 87545

lascience@lanl.gov
Tel: 505-667-1447
Fax: 505-665-4408

http://www.lanl.gov/science/

Science-Based Prediction for Complex Systems

Los Alamos
NATIONAL LABORATORY
EST. 1943

The topic of this volume, science-based prediction for complex systems, or 'predictive science' for short, is often met with questions. Hasn't science been predictive since the time of Galileo? Haven't we counted on Newton's laws to put a man on the moon and on Maxwell's equations and the constancy of Earth's gravitational field for the fantastic accuracy of the Global Positioning System? So, what's new here, and why has development of predictive capability been named as a primary technical goal of Los Alamos National Laboratory?

Although not entirely new, the pairing of prediction with complex systems makes explicit a growing expectation for accurate predictions, be they about the weather, the growth of foreign markets, or the next moves of terrorist groups. At Los Alamos, the goal is implicit in many aspects of our major missions: from predicting the reliability of our nuclear weapons without further testing to assessing the likely performance over the next 10,000 years of the proposed Yucca Mountain repository for nuclear waste and from developing strategies for detecting the smuggling of nuclear materials to inventing an optimal vaccine strategy for preventing a flu pandemic. The challenges derive not only from the complexity of the problems, but also from the degree of confidence required of the solutions and from the limited data and resources available for solving the problems.

Complex systems, as defined here, involve some combination of nonlinearity, coupled subsystems, and multiple length and time scales. These complexities invariably mean that a system can traverse many different histories, and therefore reliable prediction and accurate assessment of the uncertainties require a probabilistic approach. Also needed are the conscious coordination and integration of experiment, theory, and computer simulation.

At Los Alamos, the major driver for predictive science is, of course, the nuclear weapons program. Since the cessation of testing, the goal of the nuclear weapons program has been to predict the performance of weapons in the stockpile through direct simulation in order to anticipate problems that might arise and then develop efficient ways to fix those problems. In a penetrating analysis that opens this volume, John Pedicini and Dwight Jaeger discuss the new guidance from Washington and then outline the factors that will determine the future nuclear deterrent. What is interesting from the perspective of this volume is the emphasis on increasing predictability by creating a robust replacement for stockpile designs, one with reduced sensitivity to manufacturing and performance variables.

Whatever decisions are made on the future nuclear deterrent, methodologies are needed to predict weapons performance through simulation and to quantify levels of uncertainty. But how does one determine the uncertainties when the simulations contain a maze of errors in input data, physics models, and solution methods? The first article on uncertainty quantification introduces specific methodologies for analyzing simulation errors for multiphysics codes such as those needed for weapons performance. It also applies the methodologies to two real-world problems: estimating the errors in shock propagation problems and predicting production from an oil reservoir. The results provide a compelling case for using error models to estimate uncertainties and, in certain cases, improve the accuracy of the simulations. Using error analysis in a different application, Los Alamos researchers report a remarkable result: a factor of 10 reduction in the uncertainty in the nuclear fission cross section. That reduction is expected to translate into more accurate predictions of weapons performance and better interpretations of past nuclear tests. In the earth sciences, where data are often relatively sparse, uncertainty quantification becomes much less precise. Results reported here on ocean current stability from different ocean models show the real difficulties in predicting global climate change, and examples from volcanology illustrate the types of approximation that feed into practical decision-making.

This volume interprets predictive science in a very inclusive way, by sampling the diverse systems and new approaches being investigated at Los Alamos. The article on net-

works is a prime example, presenting a new paradigm for describing the interactions in complex systems, whether they consist of people, computers, or the complex molecules of life. The efficiency of information transport on a network seems to strongly influence the network's structural evolution, be it the Internet, the metabolic networks, or a network of scientific collaboration. That idea has led to the solution of several problems, including the design of a computer network for performing agent-based simulations in a scalable fashion.

The article on modeling the response of the retina to visual stimuli outlines another intellectual frontier. Inspired in part by the program to develop a retinal prosthetic for the visually impaired, modeling and experiment have uncovered a mechanism by which the retina may preprocess information on incoming light stimuli. What seems to be a coordinated, context-related neuronal response may also be relevant for understanding the processing that occurs deep within the brain.

Two remarkable developments are reported here on predicting material behavior under extreme conditions. One is predicting the static, dynamic, and optical properties of partially ionized matter using the framework of quantum molecular dynamics. This methodology has correctly predicted the equation of state of hydrogen and of a mixture of nitrogen and oxygen in the shocked state, as well as the viscosity of plutonium. The second development is the validation of material models that predict the deformation and fracture of metals under extreme loading conditions. The extraordinary agreement between simulation and experiment for the degree of strain localization during both tensile tests and explosively driven conditions represents the state of the art in that field.

The problem of predicting turbulence has been recalcitrant to solution for over 80 years. This volume contains an introduction to the problem through the eyes of an experimentalist followed by a discussion of exciting new developments. They include a calculation of the entire turbulent velocity field in a periodic domain, done on the Los Alamos (Advanced Simulation and Computing) Q machine. This calculation shows that the famous Kolmogorov scaling laws hold locally in time but also indicates departures. In fact, a related article on field theory and statistical hydrodynamics reports the first analytical calculation of anomalous scaling in passive scalar turbulence. Also presented is a new model for computing turbulence, known as the LANS-alpha model. Its derivation from Hamilton's principle of least action, the existence and properties of its solutions, its application to benchmark problems, its preservation of properties such as the variability of the flow, and the open problems for increasing its applicability are discussed.

The volume closes with one of the most important efforts related to the accurate simulation of nuclear weapons performance, that of developing numerical methods preserving the most important aspects of the physics. This endeavor began more than 50 years ago, at the inception of electronic digital computers. Here, in a presentation meant to be pedagogical, one gets a glimpse of the creative effort involved in making radiation and hydrodynamic simulations predictive.

All the articles reveal the impact of computational power on the progress toward predictive capability. That power is almost taken for granted, and the center of attention has shifted to what one can do with it, but it is interesting to recall that 30 years ago, when the first Cray computers were delivered to Los Alamos, computing power was less than it is today by a factor of 10^4. Most simulations were one dimensional; that is, they assumed spherical symmetry, and none of the complexity being addressed today was imagined within reach. We've come a long way.

Los Alamos Science

Science-Based Prediction for Complex Systems

About This Volume. . . . . *iv*

Weapons Outlook

The Evolving Deterrent 2
Dwight Jaeger and John Pedicini

Uncertainty Quantification

Error Analysis and Simulations of Complex Phenomena 6
Michael A. Christie, James Glimm, John W. Grove, David M. Higdon, David H. Sharp, and Merri M. Wood-Schultz

Reducing Uncertainty in Nuclear Data 26
Mark B. Chadwick, Patrick Talou, and Toshihiko Kawano

The Ocean Perspective—Uncertainties in Climate Prediction 42
Rainer Bleck

Predicting Risks in the Earth Sciences—Volcanological Examples 56
Greg Valentine

Materials

Quantum Molecular Dynamics—Simulating Warm, Dense Matter 70
Lee A. Collins, Joel D. Kress, and Stephane F. Mazevet

Predicting Material Strength, Damage, and Fracture—The Synergy between Experiment and Modeling 80
George T. (Rusty) Gray III, Paul J. Maudlin, Lawrence M. Hull, Q. Ken Zuo, and Shuh-Rong Chen

Networks

Complex Networks—The Challenge of Interaction Topology 94
Zoltán Toroczkai

Computational Biology

Models of the Retina with Application to the Design of a Visual Prosthesis 110
Garrett T. Kenyon, John George, Bryan Travis, and Krastan Blagoev

Turbulence

The Turbulence Problem—An Experimentalist's Perspective 124
Robert Ecke
Intermittency and Anomalous Scaling in Turbulence 136
Misha Chertkov

Direct Numerical Simulations of Turbulence—Data Generation and Statistical Analysis 142
Susan Kurien and Mark A. Taylor

The LANS-α Model for Computing Turbulence—Origins, Results, and Open Problems 152
Darryl D. Holm, Chris Jeffery, Susan Kurien, Daniel Livescu, Mark A. Taylor and Beth A. Wingate
Taylor's Hypothesis, Hamilton's Principle, and the LANS-α Model for Computing Turbulence. 172
Darryl D. Holm

Field Theory and Statistical Hydrodynamics—The First Analytical Predictions of Anomalous Scaling 181
Misha Chertkov

Numerical Methods

Physically Motivated Discretization Methods—A Strategy for Increased Predictiveness 188
Dana Knoll, Jim Morel, Len Margolin, and Misha Shashkov

Erratum 212

The Evolving Deterrent

Dwight Jaeger and John Pedicini

Nuclear deterrence provided the foundation of our national security strategy for the second half of the 20th century. The end of the Cold War marked the beginning of a period of transition, during which the role of nuclear weapons was uncertain. However, according to national guidance that includes the 2001 Quadrennial Defense Review, the 2002 Nuclear Posture Review, and the 2002 National Security Strategy, as well as the recommendations contained in the 2004 Defense Science Board Task Force report titled "Future Strategic Strike Forces," the direction for nuclear weapons is becoming clearer.

Synthesis of a New Direction

The overall theme of the guidance documents mentioned above is that nuclear weapons have an enduring role for a range of national security objectives, including deterrence. However, the Cold War stockpile needs to be modified to achieve U.S. defense policy goals in the 21st century. The premise of deterrence is that our adversaries believe that, if they attack the United States or our allies with weapons of mass destruction, we have the capability and,

if required, the will to destroy what they value most. To deter, we "hold at risk" those assets that are most important to an adversary. Much of the Cold War arsenal was optimized to hold at risk large nuclear forces, leadership facilities, and other important targets in large countries harboring many ready-to-deliver weapons presumably aimed at the United States. As potential adversaries have changed and nonnuclear weapons have improved, the role of nuclear deterrence has evolved toward holding at risk a much smaller number of specific targets that cannot be confidently destroyed by conventional munitions. The perceived requirements of nuclear deterrence and supporting capabilities for an unknown future are the following: Nuclear testing should not be required, collateral damage should be minimized, deterrence plans should be sufficiently flexible to meet emerging or future Department of Defense requirements, the infrastructure should be flexible and responsive if or when needed, environmental problems related to manufacturing must be minimized, cost of manufacturing and operations should be reduced, safety and security in a post 9/11 world need to be improved, and capable and well-trained stewards are necessary to ensure the continued viability of the deterrent. In our judgment, the future deterrent will likely consist of reduced numbers of existing warheads (or functional replacements for them) and the capability to build a modest number of special-capability weapons should that become necessary.

Meeting these kinds of requirements drives the physics package designers from Los Alamos and Lawrence Livermore National Laboratories and the underlying science and technology toward two goals. The first goal is to ensure that the existing systems are sustainable. Achieving this goal is currently based on life extension programs (LEPs) for most of the existing warheads. The planned LEPs are consistent with the Moscow Treaty and the recently revised (June 2004) Nuclear Warhead Stockpile Plan. Another option for achieving this goal is to develop a reliability replacement warhead (RRW)[1] to facilitate replacement of stockpile warheads and warhead components within existing requirements of the current systems. This option is now being examined at Los Alamos. The second goal is to ensure that the NNSA can provide the capabilities that may be needed to hold at risk other potential emerging types of targets, mainly deeply buried command bunkers and biological and chemical weapons, should the need for such weapons be determined by the U.S. government sometime in the future.

September 11, 2001

Ensuring the Existing Capability in the 21st Century. Whether to develop additional weapons concepts is a topic of continuing debate, but there is general consensus about the need to ensure that the existing weapon systems are sustainable. To achieve this goal, we need to rely on the underlying science and capability to predict when problems will arise. We then need the capability to replicate the parts, components, and systems in a configuration that is acceptably close to what was tested and certified. Finally, we need capable and trained people to make all this happen.

To date, our Stockpile Stewardship Program (SSP) has been quite successful. We are currently executing LEPs and considering additional LEPs for the remainder of the stockpile, perhaps on a recurring basis. This program, however, is proving to be more time-consuming and expensive than originally envisioned.

Several factors contribute to the expense of the SSP. During the Cold War, U.S. nuclear weapons were designed to meet stringent safety and security requirements while simultaneously meeting very demanding sets of military requirements; these weapons are thus highly optimized. Within a given package, enduring stockpile warheads were designed to have maximum nuclear yield (explosive power) given the highly constrained weight and volume limits of the delivery systems. These optimized, sophisticated designs left little margin for uncertainties of performance. In this context, margin is the generic term that represents the difference between where a variable operates and the upper limit capability of that variable (for example, the difference between the stress in a bridge beam at full load compared with the ultimate stress capability of the beam). Factors providing extra performance margin were secondary. Among them are the weapons' ability to perform "as designed" in a variety of adverse circumstances (for example, extreme heat or cold, radiation environments, and others), to be insensitive to small design flaws or deterioration from aging, and to be straightforward to manufacture and maintain. Considering the factors that provide extra performance margin as secondary in importance was acceptable, in part, because underground nuclear testing could be used to confirm that high-performance designs with moderate design margins would indeed work. Further, because new or replace-

[1] The RRW was recently approved for FY05 funding by the 108th Congress.

ment weapons were constantly being designed, built, and fielded to replace older weapons, age was not a significant consideration. At present, however, new parts and components must be constructed with very tight tolerances on geometry, materials, and manufacturing processes to sustain these highly optimized systems.

Los Alamos is investigating an alternate approach to ensure that the United States can maintain the existing capability through initial examination of the feasibility of an RRW. This feasibility study is concentrating on two major questions: (1) Can we certify a replacement design without nuclear testing? (2) Would such a design provide adequate or more capability with fewer resources?

In answer to the first question, we need to design replacements, bearing in mind that we must certify without nuclear testing. Such designs require development of a different set of requirements. General guidance and constraints must be defined first. A warhead must (1) be certifiable and safe, (2) meet modern surety standards and post 9/11 surety issues, (3) have larger margins with known uncertainties for all physics and engineering design variables (several standard deviations away from known failures using a formal methodology for quantification of margins and uncertainties), (4) be modular and compatible with as many delivery systems as possible, (5) have minimal susceptibility to aging changes, (6) be easier to manufacture than current warheads in the stockpile, (7) be produced for less than typical cost, have fewer parts, and be less complex, (8) whenever possible, contain fewer materials that would pose environmental risk, and (9) be field inspectable and maintainable. An RRW program would also inherently create challenging real-world environments for new stewards.

Los Alamos is building the capability to evaluate the relative costs of different scenarios for stockpile evolution. One can speculate that eight quite highly optimized warhead types (current plan) would cost more than three or four relatively simple long-life systems designed according to the criteria listed above. However, it is important to validate such assumptions before making major investments.

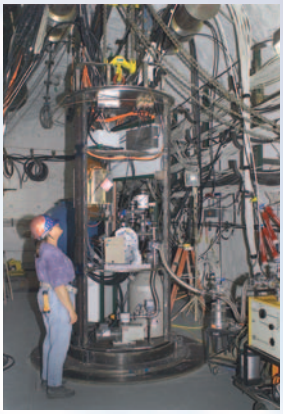
A final issue we will have to address before making a major commitment is the value of stockpile diversity. It has oftentimes been assumed that national security might be better served by a highly diverse stockpile. However, in a fixed-budget, highly constrained environment, the nation must make informed decisions about the value of many warhead types against the advantage of having a better understanding of fewer warhead types. This part of the puzzle is arguably one of the more important issues to be resolved and ultimately may be one of the hardest to address.

Providing Capabilities to Meet Future Threats. The Nuclear Posture Review also calls for the examination of nuclear weapon concepts that would be capable of neutralizing weapons of mass destruction (biological and chemical weapons) and holding at risk hard and deeply buried targets (HDBTs) that could be used to protect an enemy's leaders or key facilities.[2] Because nuclear weapons produce very high temperatures and can produce large amounts of radiation, they are lethal to biological and chemical agents. For example, some preliminary analysis indicates that neutralizing weapons of mass destruction with nuclear weapons would likely cause substantially fewer collateral casualties than might result from dispersal of biological agents under a conventional attack. However, any final assessment of potential collateral damage would require significant research.

HDBTs present a different set of challenges. A significant ground shock is required to destroy many of these types of targets. If a weapon can penetrate the ground, more of the energy is coupled directly into the ground, producing a shockwave. Typically, the effect of an underground burst can be from 20 to 50 times (depending on depth of burial) more effective than an equivalent surface burst. Stated another way, one can lower the required explosive power by the same factor. Current conventional penetrating weapons, holding less than 2000 pounds (or 1 ton) of high explosive, can hold at risk many targets buried at shallow depths. However, numerous critical targets are too deep underground and are too hard to be threatened by these systems. The United States could, in principle, develop a small number of conventional penetrators that are roughly ten times larger than current conventional bombs. These larger systems, although difficult to deliver in any numbers, could be effective at destroying some targets that are not now held at risk by nonnuclear weapons. However, adversaries could easily outdig such a capability. On the other hand, nuclear earth-penetrating weapons could be designed with a range of destructive power. This power could be adjusted to minimize collateral damage while still destroying the target. Collateral damage can be reduced through ground penetration but would produce some air shockwaves (ground shock requirements would be just high enough to destroy the target), thermal radiation, and residual dispersed radiation. However, considerable analysis of weapons' effects is required before a proposal for a warhead can be made. Pursuing these concepts beyond the idea stage is controversial, and recent legislation has removed funding for nuclear earth penetrators or advanced nuclear weapons concepts. ■

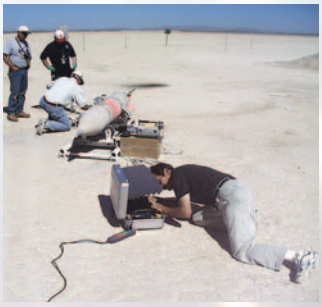
[2] Any decision to actively pursue such weapons must involve the development of Department of Defense requirements and the concurrence of Congress.

For further information, contact Dwight L. Jaeger 505 665 3797 (jaeger@lanl.gov).

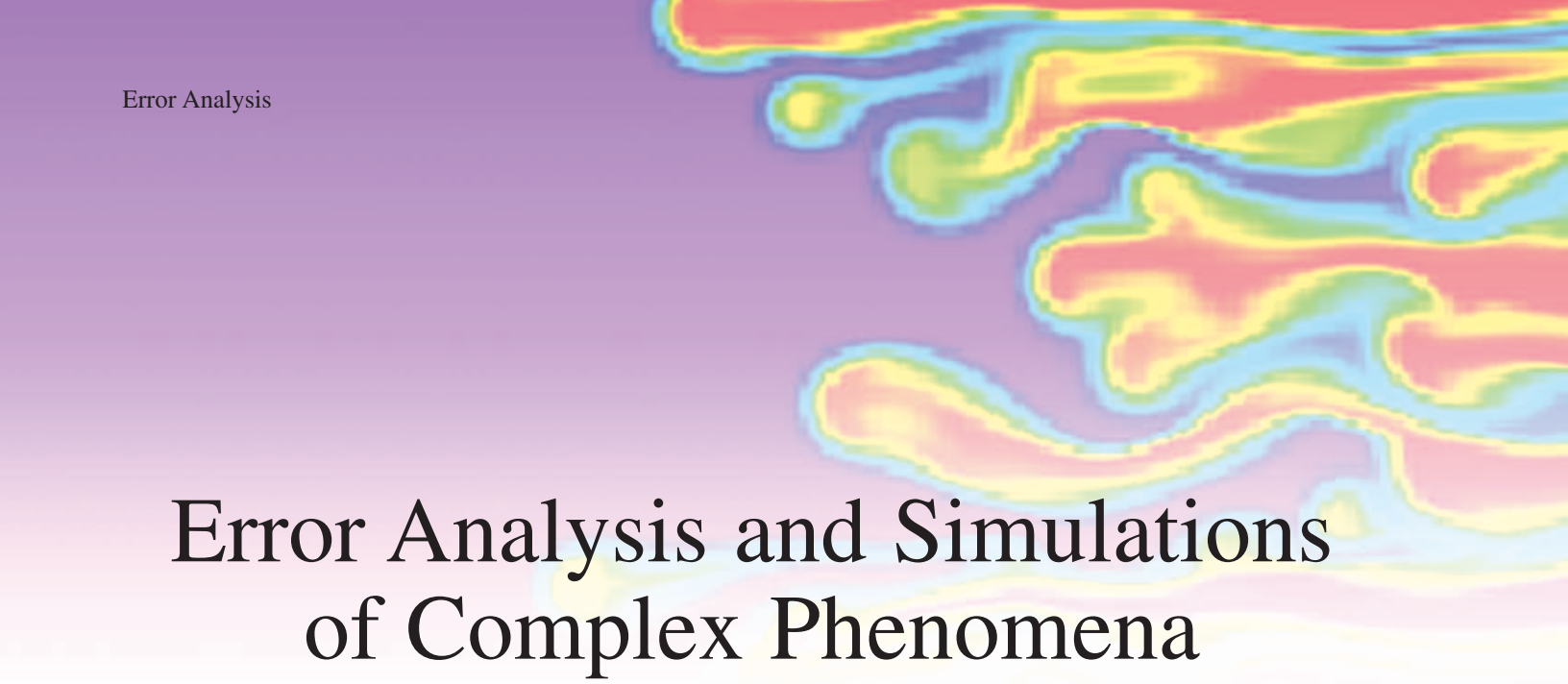


Results from subcritical experiments conducted at the Nevada Test Site are used in building predictive capabilities for stockpile certification.

Warhead disassembly and reassembly are routinely done to ensure that all systems in the stockpile are reliable. The W88 warhead at right has its reentry body wrapped in red protective material for a safer surveillance process. Los Alamos engineers and personnel from the Pantex Plant in Amarillo, Texas, improved the design of the assembly stand to enhance worker safety.



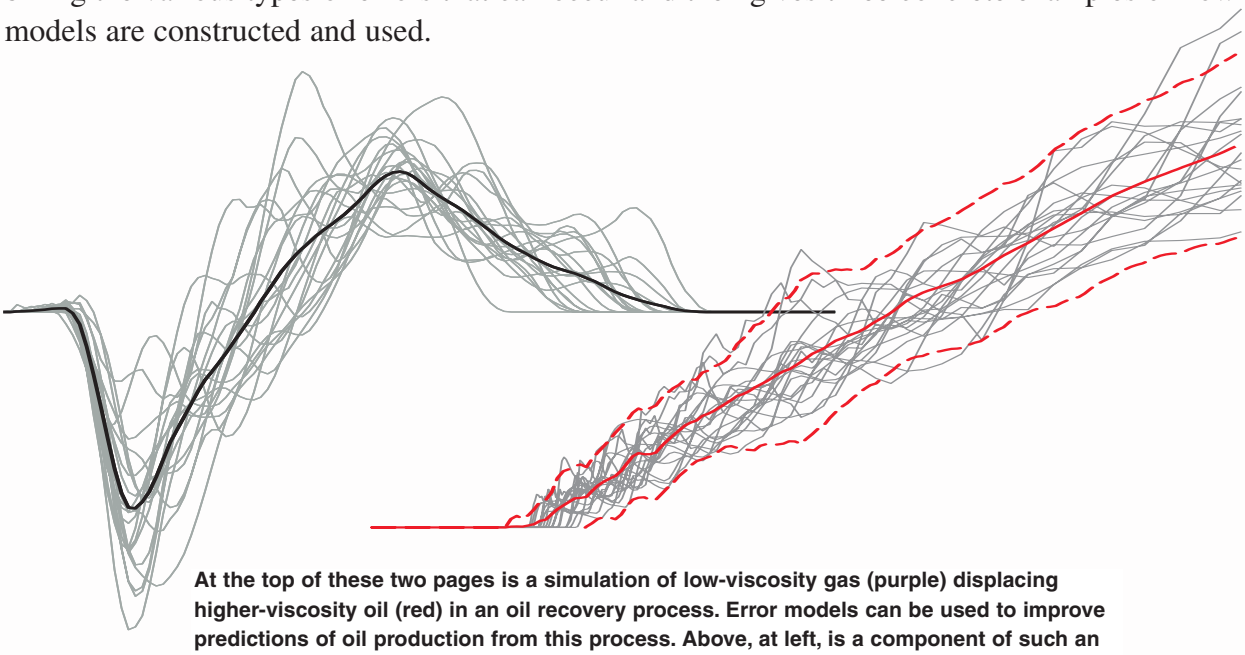
Working with other National Nuclear Security Administration and military organizations, Los Alamos staff help conduct surveillance tests, in which mockups of nuclear weapons are subjected to realistic situations to demonstrate their reliability. In this surveillance test, a B-61 look-alike weapon is dropped from a B-2 bomber (top), recovered (middle), and prepared for post-test data interrogation and radiography (bottom).



Error Analysis and Simulations of Complex Phenomena

Michael A. Christie, James Glimm, John W. Grove, David M. Higdon, David H. Sharp, and Merri M. Wood-Schultz

Large-scale computer-based simulations are being used increasingly to predict the behavior of complex systems. Prime examples include the weather, global climate change, the performance of nuclear weapons, the flow through an oil reservoir, and the performance of advanced aircraft. Simulations invariably involve theory, experimental data, and numerical modeling, all with their attendant errors. It is thus natural to ask, "Are the simulations believable?" "How does one assess the accuracy and reliability of the results?" This article lays out methodologies for analyzing and combining the various types of errors that can occur and then gives three concrete examples of how error models are constructed and used.



At the top of these two pages is a simulation of low-viscosity gas (purple) displacing higher-viscosity oil (red) in an oil recovery process. Error models can be used to improve predictions of oil production from this process. Above, at left, is a component of such an error model, and at right is a prediction of future oil production for a particular oil reservoir obtained from a simple empirical model in combination with the full error model.

Reliable Predictions of Complex Phenomena

There is an increasing demand for reliable predictions of complex phenomena encompassing, where possible, accurate predictions of full-system behavior. This requirement is driven by the needs of science itself, as in modeling of supernovae or protein interactions, and by the need for scientifically informed assessments in support of high-consequence decisions affecting the environment, national security, and health and safety. For example, decisions must be made about the amount by which greenhouse gases released into the atmosphere should be reduced, whether and for what conditions a nuclear weapon can be certified (Sharp and Wood-Schulz 2003), or whether development of an oil field is economically sound. Large-scale computer-based simulations provide the only feasible method of producing quantitative, *predictive* information about such matters, both now and for the foreseeable future. However, the cost of a mistake can be very high. It is therefore vitally important that simulation results come with a high level of confidence when used to guide high-consequence decisions.

Confidence in expectations about the behavior of real-world phenomena is typically based on repeated experience covering a range of conditions. But for the phenomena we consider here, sufficient data for high confidence is often not available for a variety of reasons. Thus, obtaining the needed data may be too hazardous or expensive, it may be forbidden as a matter of policy, as in the case of nuclear testing, or it just may not be feasible. Confidence must then be sought through understanding of the scientific foundations on which the predictions rest, including limitations on the experimental and calculational data and numerical methods used to make the prediction. This understanding must be sufficient to allow quantitative estimates of the level of accuracy and limits of applicability of the simulation, including evidence that any factors that have been ignored in making the predictions actually have a small effect on the answer. If, as sometimes happens, high-confidence predictions cannot be made, this fact must also be known, and a thorough and accurate uncertainty analysis is essential to identify measures that could reduce uncertainties to a tolerable level, or mitigate their impact.

Our goal in this paper is to provide an overview of how the accuracy and reliability of large-scale simulations of complex phenomena are assessed, and to highlight the role of what is known as an error model in this process.

Why Is It Hard to Make Accurate Predictions of Complex Phenomena?

We begin with a couple of examples that illustrate some of the uncertainties that can make accurate predictions difficult. In the oil industry, predictions of fluid flow through oil reservoirs are difficult to make with confidence because, although the fluid properties can be determined with reasonable accuracy, the fluid flow is controlled by the poorly known rock permeability and porosity. The rock properties can be measured by taking samples at wells, but these samples represent only a tiny fraction of the total reservoir volume, leading to significant uncertainties in fluid flow predictions. As an analogy of the difficulties faced in predicting fluid flow in reservoirs, imagine drawing a street map of London and then predicting traffic flows based on what you see from twelve street corners in a thick fog!

In nuclear weapons certification, a different problem arises. The physical processes in an operating nuclear weapon are not all accessible to laboratory experiments (O'Nions et al. 2002). Since underground testing is excluded by the Comprehensive Test Ban Treaty (CTBT), full system predictions can only be compared with limited archived test data.

The need for reliable predictions is not confined to the two areas above. Weather forecasting, global climate modeling, and complex engineering projects, such as aircraft design, all generate requirements for reliable, quantitative predictions—see, for example, Palmer (2000) for a study of predictability in weather and climate simulations. These often depend on features that are hard to model at the required level of detail—especially if many simulations are required in a design-test-redesign cycle.

More generally, because we are

dealing with complex phenomena, knowledge about the state of a system and the governing physical processes is often incomplete, inaccurate, or both. Furthermore, the strongly nonlinear character of many physical processes of interest can result in the dramatic amplification of even small uncertainties in the input so that they produce large uncertainties in the system behavior. The effects of this sensitivity will be exacerbated if experimental data are not available for model selection and validation. Another factor that makes prediction of complex phenomena very difficult is the need to integrate large amounts of experimental, theoretical, and computational information about a complex problem into a coherent whole. Finally, if the important physical processes couple multiple scales of length and time, very fast and very high memory capacity computers and sophisticated numerical methods are required to produce a high-fidelity simulation. The examples discussed in this article exhibit many of these difficulties, as well as the uncertainties in prediction to which they lead.

To account for such uncertainties, models of complex systems and their predictions are often formulated probabilistically. But the accuracy of predictions of complex phenomena, whether deterministic or probabilistic, varies widely in practice. For example, estimates of the amount of oil in a reservoir that is at an early stage of development are very uncertain. Large capital investments are made on the basis of probabilistic estimates of oil in place, so that the oil industry is fundamentally a risk-based business. The estimates are usually given at three confidence levels: $p90$, $p50$, and $p10$, meaning that there is a 90 percent, 50 percent, and 10 percent chance, respectively, that the amount of oil in place will be greater than the specified reserve level. Figure 1 shows a schematic plot (based on a real North Sea example) of estimated reserves as a function of time. The plot clearly shows that, as more information about the reservoir was acquired during the course of field development, estimates of the range of reserves changed out-

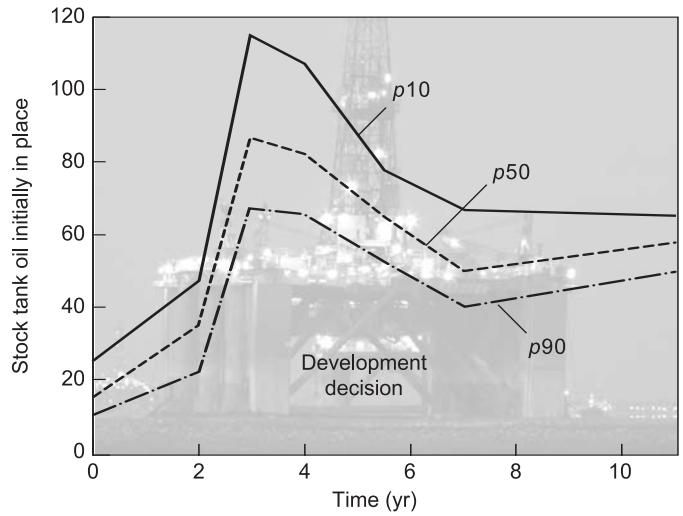


Figure 1. Oil-in-Place Uncertainty Estimate Variation with Time
This figure shows estimates of *p*90, *p*50, and *p*10 probabilities that the amount of oil in a reservoir is greater than the number shown. The estimated probabilities are plotted as a function of time. The variations shown indicate the difficulties involved in accurate probability estimations. [Photo courtesy of Terrington (York) Ltd.]

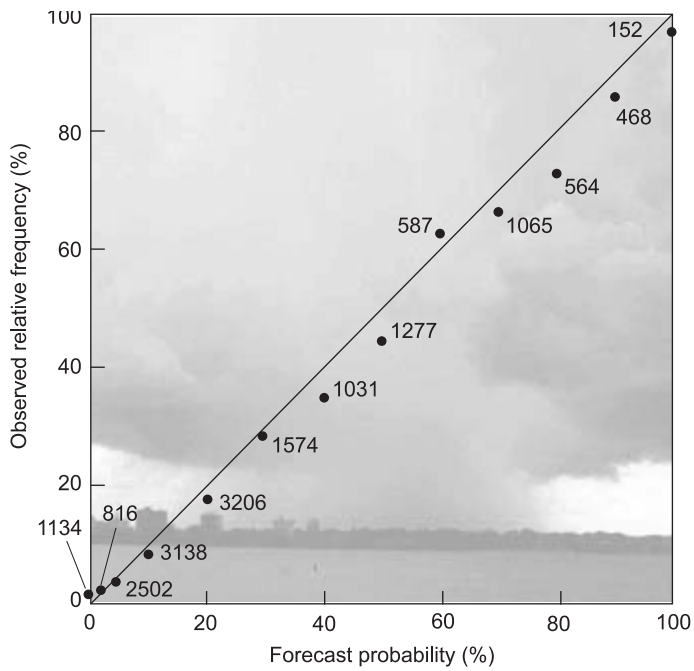


Figure 2. Calibration Curve for Weather Forecasts
This plot shows estimates of the probability of precipitation from simulation forecasts vs the observed frequency of precipitation for a large number of observations. Next to each data point is the number of observations for that forecast.

side the initial prediction. In other words, the initial estimates of reserves, although probabilistic, did not capture the full range of uncertainty and were thus unreliable. This situation was obviously a cause for concern for a company with billions of dollars in investments on the line.

Probabilistic predictions are also used in weather forecasting. If the probabilistic forecast "20 percent chance of rain" were correct, then on average it would have rained on 1 in 5 days that received that forecast. Data on whether or not it rained are easily obtained. This rapid and repeated feedback on weather predictions has resulted in significantly improved reliability of forecasts compared with predictions of uncertainty in oil reserves. The comparison between the observed frequency of precipitation and a probabilistic forecast for a locality in the United States shown in Figure 2 confirms the accuracy of the forecasts.

This accuracy did not come easily, and so we next briefly describe two of the principal methods currently used to improve the accuracy of predictions of complex phenomena: calibration and data assimilation.

Calibration is a procedure whereby a simulation is matched to a particular set of experimental data by performing a number of runs in which uncertain model parameters are varied to obtain agreement with the selected data set. This procedure is sometimes called "tuning," and in the oil industry it is known as history matching. Calibration is useful when codes are to be used for interpolation, but it is of limited help for extrapolation outside the data set that was used for tuning. One reason for this lack of predictability is that calibration only ensures that unknown errors from different sources, say inaccurate physics and numerics, have been adjusted to compensate one another, so that the net error in some observable is small. Because different physical processes and numerical errors are unlikely to scale in the same way, a calibrated simulation is reliable only for the regime for which it has been shown to match experimental data.

In one variant of calibration, multiple simultaneous simulations are performed with different models. The "best" prediction is defined as a weighted average over the results obtained with the different models. As additional observations become available, the more successful models are revealed, and their predictions are weighted more heavily. If the models used reflect the range of modeling uncertainty, then the range of results will indicate the variance of the prediction due to those uncertainties.

Data assimilation, while basically a form of calibration, has important distinctive features. One of the most important is that it enables real-time utilization of data to improve predictions. The need for this capability comes from the fact that, in operational weather forecasting, for example, there is insufficient time to restart a run from the beginning with new data, so that this information must be incorporated on the fly. In data assimilation, one makes repeated corrections to model parameters during a single run, to bring the code output into agreement with the latest data. The corrections are typically determined using a time series analysis of the discrepancies between the simulation and the current observations. Data assimilation is widely used in weather forecasting. See Kao et al. (2004) for a recent application to shock-wave dynamics.

Sources of Error and How to Analyze Them

Introducing Error Models. The role of a thorough error analysis in establishing confidence in predictions has been mentioned. But evaluating the error in a prediction is often more difficult than making the prediction in the first place, and when confidence in the answer is an issue, it is just as important.

A systematic approach for determining and managing error in simulations is to try to represent the effects of inaccurate models, neglected phenomena, and limited solution accuracy using an error model.

Unlike the calibration and data assimilation methods discussed above, an error model is not primarily a method of increasing the accuracy of a simulation. Error modeling aims to provide an independent estimate of the *known* inadequacies in the simulation. An error model does not purport to provide a complete and precise explanation of observed discrepancies between simulation and experiment or, more generally, of the differences between the simulation model and the real world. In practice, an error model helps one achieve a scientific understanding of the knowable sources of error in the simulation and put quantitative bounds on as much of the error as possible.

Simulation Errors. Computer codes used for calculating complex phenomena combine models for diverse physical processes with algorithms for solving the governing equations. Large databases containing material properties such as cross sections or equations of state that tie the simulation to a real-world system must be integrated into the simulation at the lowest level of aggregation. These components and, significantly, input from the user of the code must be linked by a sophisticated computer science infrastructure, with the result that a simulation code for complex phenomena is an exceedingly elaborate piece of software. Such codes, while elaborate, still provide only an approximate representation of reality.

Simulation errors come from three

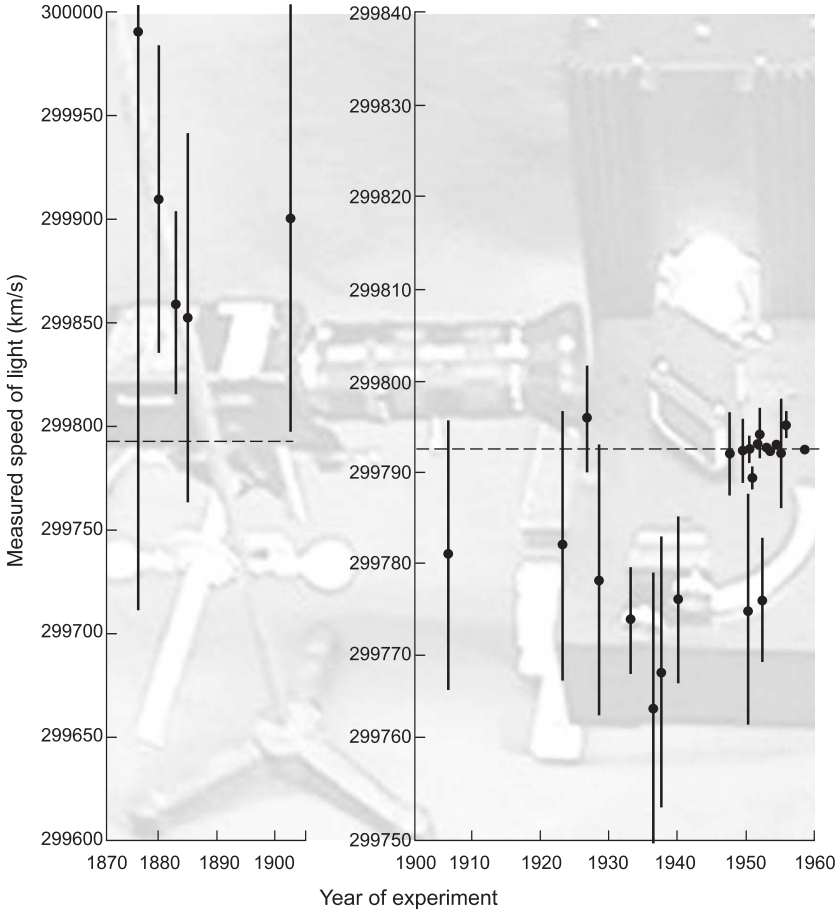


Figure 3. Uncertainties in Reported Measurements of the Speed of Light (1870–1960)

This figure shows measured values of the speed of light along with estimates of the uncertainties in the measured values up until 1960. The error bars correspond to the estimated 90% confidence intervals. The currently accepted value lies outside the error bars of more measurements than would be expected, indicating the difficulty of truly assessing the uncertainty in an experimental measurement. Refer to the article by Henrion and Fischoff on pp. 666–677 in *Heuristics and Biases* (2002) for more details on this and other examples of uncertainties in physical constants.

(Photo courtesy of Department of Physics, Carnegie Mellon University.)

main sources: inaccurate input data, inaccurate physics models, and limited accuracy of the solutions of the governing equations. Clearly, each of these generic sources of error is potentially important. A perfect physics model with perfect input data will give wrong answers if the equations are solved poorly. Likewise, a perfect solution of the wrong equations will also give incorrect answers. The relative importance of errors from each source is problem dependent, but each source of error must be evaluated. Our discussion of error models will reflect the above comments by categorizing simulation inadequacies as due to input, solution, and physics errors.

Input errors refer to errors in data used to specify the problem, and they include errors in material properties, the description of geometrical configurations, boundary and initial conditions, and others. Solution error is the difference between the exact mathematical solution of the governing equations for the model and the approximate solution of the equations obtained with the numerical algorithms used in the simulation. Physics error includes the effects of phenomena that are inadequately represented in the simulation, for example, the unknown details of subscale physics, such as the microscopic details of material that is treated macroscopically in the simulation. Evaluations of the effects of these details are typically based on statistical descriptions. The physics component of an error model is thus based on knowledge of aspects of the nominal model that need or might need correction.

Experimental Errors and Solution Errors. Much of our understanding of how to analyze errors comes from studies of experimental error. We will also see below that experimental and solution errors play a similar role in an uncertainty analysis. We therefore start by discussing experimental errors.

Experimental errors play an important role in building error models for simulations. First, they can bias conclusions that are drawn when simulation results are compared with measured data. Second, experimental errors affect the accuracy of simulations indirectly through their effects on databases and input data used in a simulation. Experimental errors are classified as random or systematic. Typically, both types of error are present in any particular application. A familiar example of a random error is the statistical sampling error quoted along with the results of opinion polls. Another type of random error is the result of variations in random physical processes, such as the number of radioactive decays in a sample per unit time. The signals from measuring instruments usually contain a compo-

nent that either is or appears to be random whether the process that is the subject of the measurement is random or not. This component is the ubiquitous "noise" that arises from a wide variety of unwanted or uncharacterized processes occurring in the measurement apparatus. The way in which noise affects a measurement must be taken into consideration to attain valid conclusions based on that data. Noise is typically treated probabilistically, either separately or included with a statistical treatment of other random error. However, systematic error is often both more important and more difficult to deal with than random error. It is also frequently overlooked, or even ignored.

To see how a systematic error can occur, imagine that an opinion poll on the importance of education was conducted by questioning people on street corners "at random"—not knowing that many of them were coming and going from a major library that happened to be located nearby. It is virtually certain that those questioned would on average place a higher importance on education than the population in general. Even if a very large number of those individuals were questioned, an activity that would result in a small statistical sampling error, conclusions about the importance of higher education drawn from this data could be incorrect for the population at large. This is why carefully conducted polls seek to avoid systematic errors, or biases, by ensuring that the population sampled is representative.

As a second example, suppose that 10 measurements of the distance from the Earth to the Sun gave a mean value of 95,000,000 miles due, say, to flaws in an electric cable used in making these measurements. How would someone know that 95,000,000 miles is the wrong answer? This error could not be revealed by a statistical analysis of only those 10 measurements. Additional, independent measurements made with independent measuring equipment would suggest that something was wrong if they were inconsistent with these results. However, the cause of the systematic error could only be identified through a physical understanding of how the instruments work, including an analysis of the experimental procedures and the experimental environment. In this example, the additional measurements should show that the electrical characteristics of the cable were not as expected. To reiterate, the point of both examples is that an understanding of the systematic error in a measured quantity requires an analysis that is independent of the instrument used for the original measurement.

An example of how difficult it can be to determine uncertainties correctly is shown in Figure 3, a plot of estimates of the speed of light vs the date of the measurement. The dotted line shows the accepted value, and the published experimental uncertainties are shown as error bars. The length of the error bars—1.48 times the standard deviation—is the "90 percent confidence interval" for a normally distributed uncertainty for the experimental error; that is, the experimental error bars will include the correct value 90 percent of the time if the uncertainty were assessed correctly. It is evident from the figure, however, that many of the analyses were inadequate: The true value lies outside the error bars far more often than 10 percent of the time. This situation is not uncommon, and it provides an example of the degree of caution appropriate when using experimental results.

The analysis of experimental error is often quite arduous, and the rigor with which it is done varies in practice, depending on the importance of the result, the accuracy required, whether the measurement technique is standard or novel, and whether the result is controversial. Often, the best way to judge the adequacy of an analysis of uncertainty in a complex experiment is to repeat the experiment with an independent method and an independent team.

Solution errors enter an analysis of simulation error in several ways. In addition to being a direct source of error in predictions made with a given model, solution errors can bias the conclusions one draws from comparing a model to data in exactly the same way that experimental errors do. Solution errors also can affect a simulation almost covertly: It is common for the data or the code output to need further processing before the two can be directly compared. When this processing requires modeling or simulation with a different code, then the solution error from that calculation can affect the comparison. As with experimental errors, solution errors must be determined independently of the simulations that are being used for prediction.

Using Data to Constrain Models

The scientific method uses a cycle of comparison of model results with data, alternating with model modification. A thorough and accurate error analysis is necessary to validate improvements. The availability of data is a significant issue for complex systems, and data limitations permeate efforts to improve simulation-based predictions. It is therefore important to use all relevant data in spite of differences in experiment design and measurement technique. This means that it is important to have a procedure to combine data from diverse sources and to understand the significance of the various errors that are responsible for limitations on predictability.

The way in which the various categories of error can affect comparison with experimental data and the steps

to be taken if the errors are too large are discussed in the next section. The comparison of model predictions with experimental data is often called the forward step in this cycle and is a key component in uncertainty assessments of a predicted result. The backward step of the cycle for model improvement, which is discussed next, is the statistical inference of an improved model from the experimental data. The Bayesian framework provides a systematic procedure for inference of an improved model from observations; lastly, we describe the use of hierarchical Bayesian models to integrate data from many sources.

Some discussion of the use of the terms "uncertainty" and "error" is in order. In general, any physical quantity, whether random or not, has a specific value—such as the number of radioactive decays in a sample of tritiated paint in a given 5-minute period. The difference between that actual number and an estimate determined from knowledge of the number of tritium nuclei present and the tritium lifetime is the error in that estimate. If the experiment were repeated many times, a distribution of errors would arise, and the probability density function for those errors is the uncertainty in the estimate.

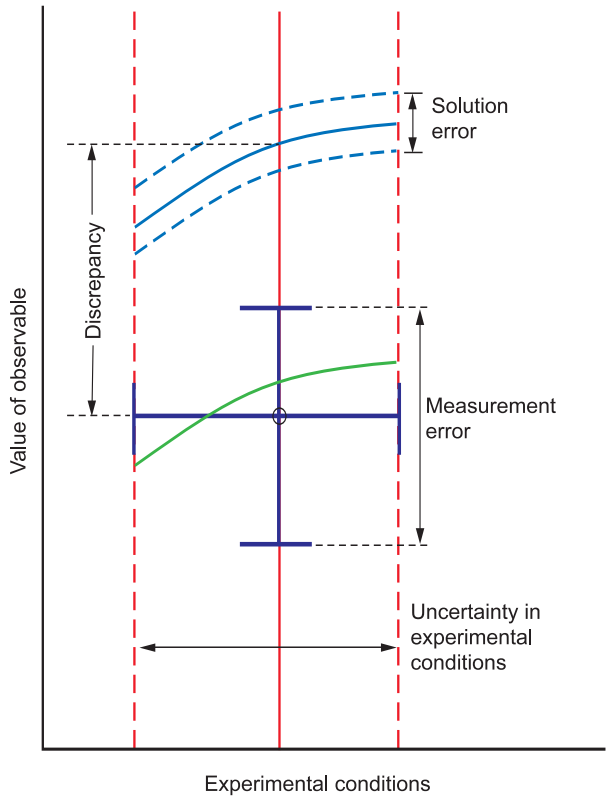


Figure 4. Comparing Experimental Measurements with Simulations
The green line shows the true, unknown value of an observable over the range of uncertainty in the experimental conditions, and the purple cross indicates the uncertainty in the observation. The discrepancy measures the difference between observation and simulation.

Decomposition of Errors. Our ability to predict any physical phenomenon is determined by the accuracy of our input data and our modeling approach. When the modeling input data are obtained by analysis of experiments, the experimental error and modeling error (solution error plus physics approximations) terms control the accuracy of our estimation of those data, and hence our ability to predict. Because a full uncertainty-quantification study is in itself a complex process, it is important to ensure that those errors whose size can be controlled—either by experimental technique or by modeling/simulation choices—are small enough to ensure that predictions of the phenomena of interest can be made with sufficient precision for the task at hand. This means that simpler techniques are often appropriate at the start of a study to ensure that we are operating with the required level of precision.

The discrepancy between simulation results and experimental data is illustrated in Figure 4, which shows the way in which this discrepancy can be related to measurement errors and solution errors. Note that the experimental conditions are also subject to uncertainties. This means that the observed value may be associated with a slightly different condition than the one for which the experiment was designed, as shown in Figure 4.

The three steps below could serve as an initial, deterministic assessment of the discrepancy between simulation and experiment.

Step 1. Compare Simulated and Experimental Results. The size of the measurement error will obviously affect the conclusions drawn from the comparison. Those conclusions can also be affected by the degree of knowledge of the actual as opposed to the designed experimental conditions. For example, the as-built composition of the physical parts of the system

under investigation may differ slightly from the original design. The effects of both of these errors are typically reported together, but they are explicitly separated here because error in the experimental conditions affects the simulated result, as well as the measured result, as can be seen in Figure 4.

Step 2. Evaluate Solution Errors. If the error is a simple matter of numerical accuracy—for example, spatial or temporal resolution—then the error is a fixed, determinable number *in principle*. In other cases—for example, subgrid stochastic processes—the error may be knowable in only a statistical sense.

Step 3. Determine Impact on Predictability. If the discrepancy is large compared with the solution error and experimental uncertainty, then the model must be improved. If not, the model may be correct, but in either case, the data can be used to define a range of modeling parameters that is consistent with the observations. If that range leads to an uncertainty in prediction that is too large for the decision being taken, the experimental errors or solution errors must be reduced.

A significant discrepancy in step 1 indicates the presence of errors in the simulation and/or experiment, and steps 2 and 3 are necessary, but not sufficient, to pinpoint the source(s) of error. However, these simple steps do not capture the true complexity of analyzing input or modeling errors. In practice, the system must be subdivided into pieces for which the errors can be isolated (see below) and independently determined. The different errors must then be carefully recombined to determine the uncertainties in integral quantities, such as the yield of a nuclear weapon or the production of an oil well, that are measured in full system tests. A potential drawback of this paradigm is that experiments on subsystems may not be able to probe the entire parameter space encountered in full system operation. Nevertheless, because the need to predict integral quantities motivates the development and use of simulation, a crucial test of the "correctness" of a simulation is that it consistently and accurately matches all available data.

Statistical Prediction

A major challenge of statistical prediction is assessing the uncertainty in a predicted result. Given a simulation model, this problem reduces to the propagation of errors from the simulation input to the simulated result. One major problem in examining the impact of uncertainties in input data on simulation results is the "curse of dimensionality." If the problem is described by a large number of input parameters and the response surface is anything other than a smooth quasilinear function of the input variables, computing the shape of the response surface can be intractable even with large parallel machines. For example, if we have identified 8 critical parameters in a specific problem and can afford to run 1 million simulations, we can resolve the response surface to an accuracy of fewer than 7 equally spaced points per axis.

Various methods exist to assess the most important input parameters. Sensitivities to partial derivatives can be computed either numerically or through adjoint methods. Adjoint methods allow computation of sensitivities in a reasonable time and are widely used.

Experimental design techniques can be used to improve efficiency. Here, the response surface is assumed to be a simple low-order polynomial in the input variables, and then statistical techniques are used to extract the maximum amount of information for a given number of runs. Principal component analysis can also be used to find combinations of parameters that capture most of the variability.

The principle that underlies many of these techniques is that, for a complex engineering system to be reliable, it should not depend sensitively on the values of, for example, 10^4 or more parameters. This is as true for a weapon system that is required to operate reliably as it is for an oil field that is developed with billions of dollars of investment funds.

Statistical Inference—The Bayesian Framework. The Bayesian framework for statistical inference provides a systematic procedure for updating current knowledge of a system on the basis of new information. In engineering and natural science applications, we represent the system by a simulation model m, which is intended to be a complete specification of all information needed to solve a given problem. Thus m includes the governing evolution equations (typically, partial differential equations) for the physical model, initial and boundary conditions, and various model parameters, but it would not generally include the parameters used to specify the numerical solution procedure itself. Any or all of the information in m may be uncertain to some degree. To represent the uncertainty that may be present in the initial specification of the system, we introduce an ensemble of models $\mathcal{M}$, with $m \in \mathcal{M}$, and define a probability distribution on $\mathcal{M}$. This is called the prior distribution and is denoted by $p(m)$.

If additional information about the system is supplied by an observation $\mathcal{O}$, one can determine an updated estimate of the probability for m, called the posterior distribution and denoted by $p(m|\mathcal{O})$, by using Bayes' formula

$$p(m|\mathcal{O}) = \frac{p(\mathcal{O}|m)\,p(m)}{\int_{\mathcal{M}} p(\mathcal{O}|m)\,p(m)\,dm} \quad . \qquad (1)$$

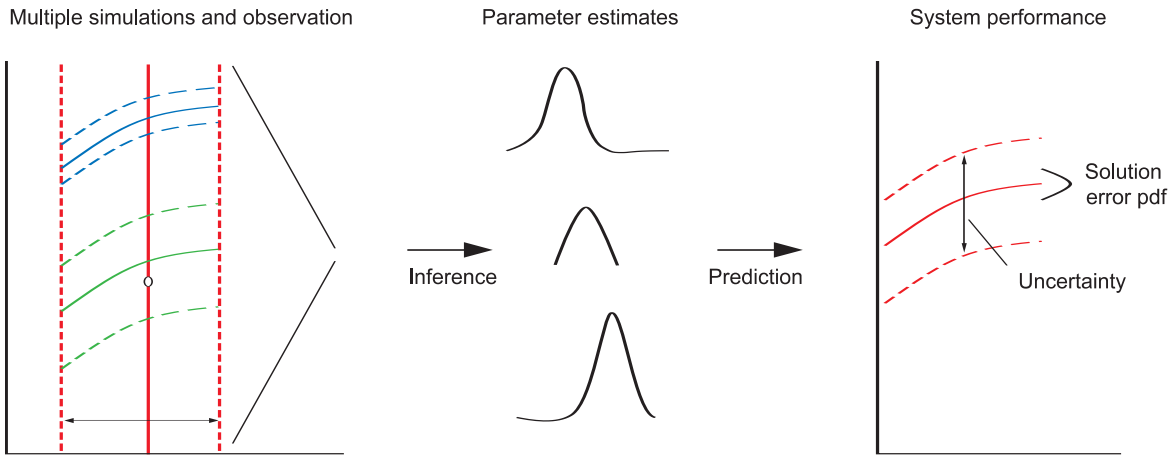


Figure 5. Bayesian Framework for Predicting System Performance with Relevant Uncertainties
Multiple simulations are performed using the full physical range of parameters. The discrepancies between the observation and the simulated values are used in a statistical inference procedure to update estimates of modeling and input uncertainties. The update involves computing the likelihood of the model parameters by using Bayes' theorem. The likelihood is computed from a probability model for the discrepancy, taking into account the measurement errors (shown schematically by the green dotted lines) and the solution errors (blue dotted lines). The updated parameter values are then used to predict system performance, and a decision is taken on whether the accuracy of the predictions is adequate.

It is important to realize that the Bayesian procedure does not determine the choice of $p(m)$. Thus, in using Bayesian analysis, one must supply the prior from an independent data source or a more fundamental theory, or otherwise, one must use a noninformative "flat" prior.

The factor $p(\mathcal{O}|m)$ in Equation (1) is called the likelihood. The likelihood is the (unnormalized) conditional probability for the observation $\mathcal{O}$, given the model m. In the cases of interest here, model predictions are determined by solutions $s(m)$ of the governing equations. The simulated observables are functionals $\mathcal{O}(s(m))$ of $s(m)$. If both the experimentally measured observables $\mathcal{O}$ and the solution $s(m)$, hence $\mathcal{O}(s(m))$, are exact, the likelihood $p(\mathcal{O}|m)$ is a delta function concentrated on the hypersurface in $\mathcal{M}$ defined by the equation

$$\mathcal{O} = \mathcal{O}\big(s(m)\big) \quad . \tag{2}$$

Real-world observations and simulations contain errors, of course, so that a discrepancy will invariably be observed between $\mathcal{O}$ and $\mathcal{O}(s(m))$. Because the likelihood is evaluated subject to the hypothesis that the model $m \in \mathcal{M}$ is correct, any such discrepancy can be attributed to errors either in the solution or in the measurements. The likelihood is defined by assigning probabilities to solution and/or measurement errors of different sizes. The required probability models for both types of errors must be supplied by an independent analysis.

This discussion shows that the role of the likelihood in simulation-based prediction is to assign a weight to a model m based on a probabilistic measure of the quality of the fit of the model predictions to data. Probability models for solution and measurement errors play a similar role in determining the likelihood.

This point is so fundamental and sufficiently removed from common approaches to error analysis that we repeat it for emphasis: *Numerical and observation errors are the leading terms in the determination of the Bayesian likelihood*. They supply critical information needed for uncertainty quantification.

Alternative approaches to inference include the use of interval analysis, possibility theory, fuzzy sets, theories of evidence, and others. We do not survey these alternatives here, but simply mention that they are based on different assumptions about what is known and what can be concluded. For example, interval analysis assumes that unknown

parameters vary within an interval (known exactly), but that the distribution of possible values of the parameter within the interval is not known even in a probabilistic sense. This method yields error bars but not confidence intervals.

An illustration of the Bayesian framework we follow to compute the impact of solution error and experimental uncertainty is shown in Figure 5. Multiple simulations are performed with the full physical range of parameters. The discrepancies (between simulation and observation) are used in a statistical inference procedure to update estimates of modeling and input uncertainties. These updated values are then used to predict system performance, and a decision is taken on whether the accuracy of the predictions is adequate.

Combining Information from Diverse Sources

Bayesian inference can be extended to include multiple sources of information about the details of a physical process that is being simulated (Gaver 1992). This information may come from "off-line" experiments on separate components of the simulation model m, expert judgment, measurements of the actual physical process being simulated, and measurements of a physical process that is related, but not identical, to the process being simulated. Such information can be incorporated into the inference process by using Bayesian hierarchical models, which can account for the nature and strength of these various sources of information. This capability is very important since data directly bearing on the process being modeled is often in short supply and expensive to acquire. Therefore, it is essential to make full use of all possible sources of information—even those that provide only indirect information.

In principle, an analysis can utilize any experimental data that can be compared with some part of the output of a simulation. To understand this point, let us make the simple and often useful assumption that the family of possible models $\mathcal{M}$ can be indexed by a set of parameters . In this case, the somewhat abstract specification of the prior as a probability distribution $p(m)$ on models can be thought of simply as a probability distribution $p(\theta)$ on the parameters θ. Depending on the application, θ may include parameters that describe the physical properties of a system, such as its equation of state, or that specify the initial and boundary conditions for the system, to mention just a few examples. In any of these cases, uncertainty in θ affects prediction uncertainty. Typically, different data sources will give information about different parameters.

Multiple sources of experimental data can be included in a Bayesian analysis by generalizing the likelihood term. If, for example, the experimental observations $\mathcal{O}$ decompose into three components ($\mathcal{O}_1$, $\mathcal{O}_2$, $\mathcal{O}_3$), the likelihood can be written as

$$p\left(\mathcal{O}\middle|m(\theta)\right) = p\left(\mathcal{O}_1\middle|m_1(\theta)\right) \times p\left(\mathcal{O}_2\middle|m_2(\theta)\right) \times p\left(\mathcal{O}_3\middle|m_3(\theta)\right)$$

if we assume that each component of the data gives information about an independent parameter θ. The subscripts on the models are there to remind us that, although the same simulation model is used for each of the likelihood components, different subroutines within the simulation code are likely to be used to simulate the different components of the output. This means that each of the likelihood terms will have its own solution error, as well as its own observation error. The relative sizes of these errors greatly affect how these various data sources constrain θ. For example, if it is known that $m_2(\theta)$ does not reliably simulate $\mathcal{O}_2$, then the likelihood should reflect this fact. Note that a danger here is that a misspecification of a likelihood term may give some data sources undue influence in constraining possible values of one of the parameters θ.

In some cases, one (or more) component (components) of the observed data is (are) not from the actual physical system of interest, but from a related system. In such cases, Bayesian hierarchical models can be used to borrow strength from that data by specifying a prior model that incorporates information from the different systems. See Johnson et al. (2003) for an example.

Finally, expert judgment usually plays a significant role in the formulation and use of models of complex phenomena—whether or not the models are probabilistic. Sometimes, expert judgment is exercised in an indirect way, through selection of a likelihood model or through the choice of the data sources to be included in an analysis. Expert judgment is also used to help with the choice of the probability distribution for $p(\theta)$, or to constrain the range of possible outcomes in an experiment, and such information is often invoked in applications for which experimental or observational data are scarce or nonexistent. However, the use of expert judgment is fraught with its own set of difficulties. For example, the choice of a prior can leave a strong "imprint" on results inferred from subsequent experiments. See *Heuristics and Biases* (2002) for enlightening discussions of this topic.

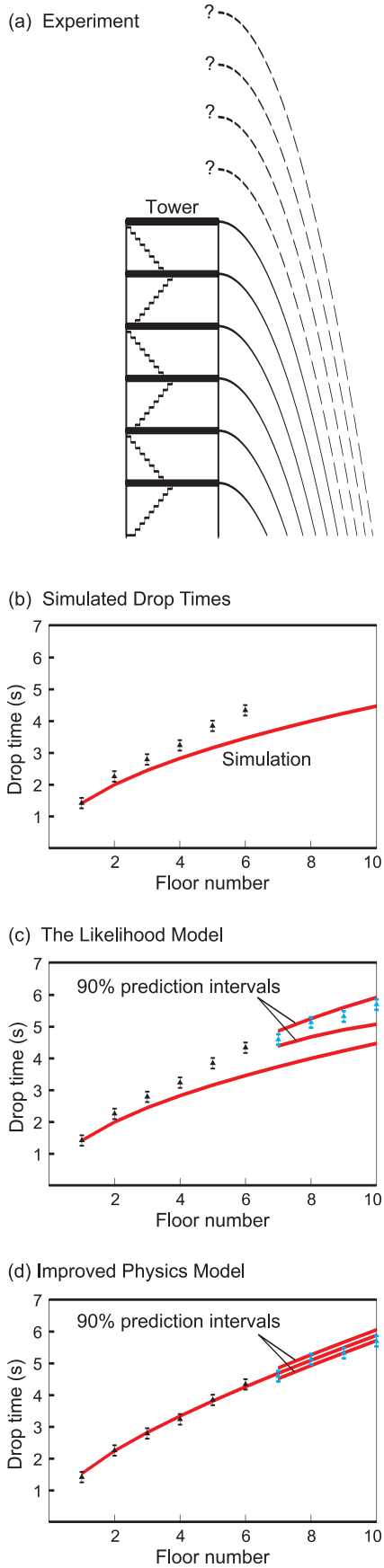


Figure 6. Dropping an Object from a Tower

(a) The time it takes an object to drop from each of 6 floors of a tower is recorded. There is an uncertainty in the measured drop times of about ±0.2 s. Predictions for times are desired for drops from floors 7 through 10, but they do not yet exist.

(b) A mathematical model is developed to predict the drop times as a function of drop height. The simulated drop times (red line) are systematically too low when compared with the experimental data (triangles). The error bars around the observed drop times show the observation uncertainty.

(c) This systematic deviation between the mathematical model and the experimental data is accounted for in the likelihood model. A fitted correction term adjusts the model-based predictions to better match the data. The resulting 90% prediction intervals for floors 7 through 10 are shown in this figure. Note that the prediction intervals become wider as the drop level moves away from the floors with experimental data. The cyan triangles corresponding to floors 7 through 10 show experimental observations taken later only for validation of the predictions.

(d) An improved simulation model was constructed that accounts for air resistance. A parameter controlling the strength of the resistance must be estimated from the data, resulting in some prediction uncertainty (90% prediction intervals are shown for floors 7 through 10). The improved model captures more of the physics, giving reduced prediction uncertainty.

Building Error Models—Examples

Dropping Objects from a Tower. Some of the basic ideas used in building error models are illustrated in Figure 6. In this example, experimental observations are combined with a simple physics model to predict how long it takes an object to fall to the ground when it is dropped from a tower. The experimental data are drop times recorded when the object is dropped from each of six floors of the tower. The actual drop time is measured with an observation error, which we assume for illustrative purposes to be Gaussian (normal), with mean 0 and a standard deviation of 0.2 second. The physics model is based solely on the acceleration due to gravity. We observe that the predicted drop times are too short and that this discrepancy apparently grows with the height from which the object is dropped.

Even though this model shows a substantial error, which is apparent from the discrepancy between the experimental data and the model predictions (Figure 6(b)), it can still be made useful for predicting drop times from heights that are greater than the height of the tower. As a first step, we account for the discrepancy by including an additional unknown correction in the initial specification of the model, namely, in the prior. This term represents the discrepancy as an unknown, smooth function of drop height that is estimated (with uncertainty) in the analysis. The results are applied to give predictions of drop times for heights that would correspond to the seventh through tenth floors of the tower. These predictions have a fair amount of uncertainty because the discrepancy term has to be extrapolated to drop heights that are beyond the range of the experimental data. Note also that the prediction uncertainty increases with drop height (refer to Figure 6(c)).

This strictly phenomenological

modeling of the error leads to results that can be extrapolated over a very limited range only, because predictions of drop times from just a few floors above the sixth have unacceptably large uncertainties. But an improved physics model can greatly extend the range over which useful predictions can be made. Thus, we next carry out an analysis using a model that incorporates a physically motivated term for air resistance. This model requires estimation of a single additional parameter appearing as a coefficient in the air resistance term. But when this parameter is constrained by experimental data, much better agreement with the measured drop times is obtained (see Figure 6(d)). In fact, in this case, the discrepancy is estimated to be nearly zero. The remaining uncertainty in this improved prediction results from uncertainties in the measured data and in the value of the air resistance parameter.

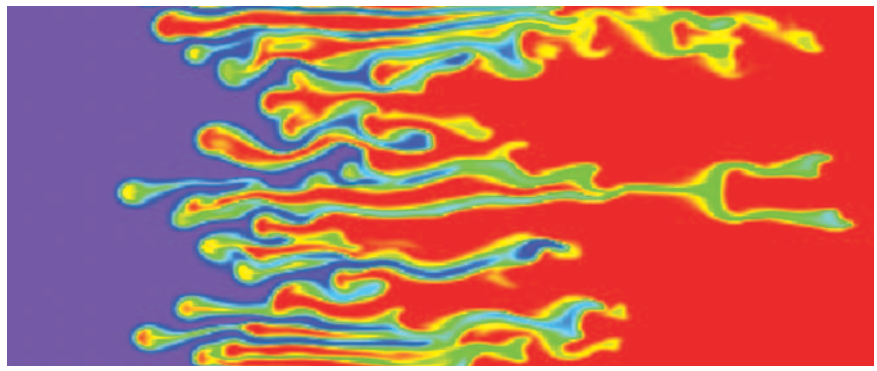


Figure 7. Viscous Fingering in a Realization of Porous Media Flow
Low-viscosity gas (purple) is injected into a reservoir to displace higher-viscosity oil (red). The displacement is unstable and the gas fingers into the oil, reducing recovery efficiency.

Using an Error Model to Improve Predictions of Oil Production. In most oil reservoirs, the oil is recovered by injecting a fluid to displace the oil toward the production wells. The efficiency of the oil recovery depends, in part, on the physical properties of the displacing fluid. The example in this section concerns estimation of the viscosity (typically poorly known) of an injected gas displacing oil in a porous medium. We will show how an error model for such estimates allows improved estimates of the uncertainty in future oil production using this method of recovery.

Because the injected gas has lower viscosity than the oil, the displacement process is unstable and viscous fingers develop (see Figure 7). The phenomenon is similar to the Rayleigh-Taylor instability of a dense fluid on top of a less dense fluid. The fingers have a reasonably predictable average behavior, but there is some randomness in their formation and evolution associated with the lack of knowledge of the initial conditions and with unknown small-scale fluctuations in rock properties.

The oil industry has a simple empirical model that accounts for the effects of fingering. This model, called the Todd and Longstaff model, fits an expansion wave (rarefaction fan) to the average behavior. Although the model is good, it is not perfect, and in particular, when applied to cases with a correlated permeability field, it tends to underestimate the speed with which the leading edge of the gas moves through the medium. If we compare results from the Todd and Longstaff model with observed data in order to estimate physical parameters such as viscosity, we will introduce errors into the parameter estimates because of the errors in the solution method. To compensate for these errors, we create a statistical model for the solution errors.[1]

For this example, we assume that the primary unknown in the Todd and Longstaff model is the ratio of gas viscosity to oil viscosity, which determines the rate at which instabilities grow. This ratio will be determined by comparing simulation and observation (in practice, oil and gas viscosities would be measured, although there would still be uncertainties associated with amounts of gas dissolved in the oil). To construct a solution error model for the average gas concentration in the reservoir, we run a number of fine-grid simulations at discrete values of the viscosity ratio, which we refer to as calibration points. Then, for each value of the viscosity ratio, we compute the difference between the Todd and Longstaff model and the fine-grid simulations as a function of scaled distance along the flow (x) and dimensionless time (t) (time divided by the time for gas to break through in the absence of fingering). The mean error computed in this way for the viscosity ratio 10 is shown in Figure 8 as a function of the similarity variable x/t. We also compute the standard deviation of the error at each time, as well as the correlation between errors at different times. This information is represented as a "covariance matrix."

We will show that the solution error model (the mean error and the covariance matrix), when used in conjunction with predictions of the Todd and Longstaff model at different viscosity ratios, can yield good estimates of the viscosity ratio for a given production data set. Figure 9 shows the

[1] All the results cited in this section are from Alannah O'Sullivan's Ph.D. thesis on error modeling (O'Sullivan 2004). We are grateful to her for permission to use these unpublished results in this article.

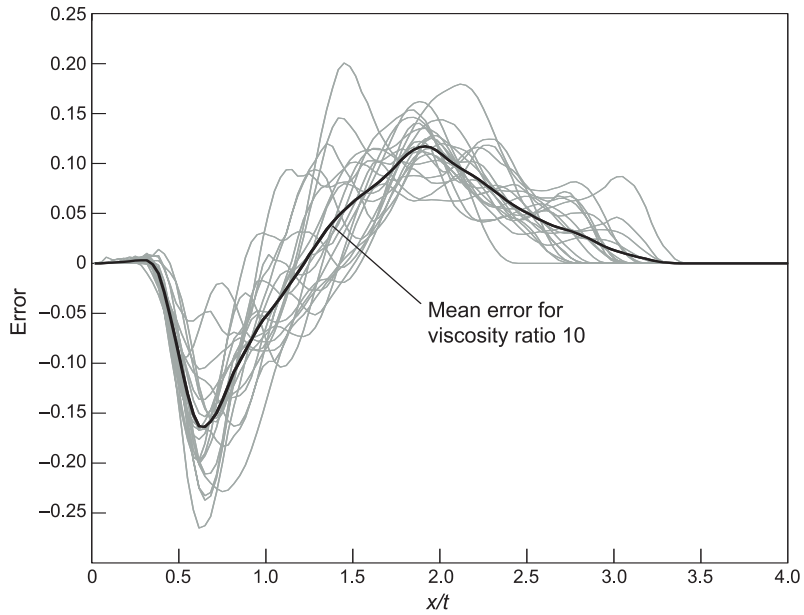


Figure 8. Mean Error and Data to Compute Mean Error
The black curve is the mean error in the gas concentration for viscosity ratio 10. The data to compute the mean error (gray curves) come from the differences between a single coarse-grid or approximate solution (in this case, the Todd and Longstaff model) and multiple fine-grid realizations, all computed at viscosity ratio 10. The variability in the fine-grid realizations reflects random fluctuations in the permeability field, which create different finger locations and growth paths. In this case, the gas concentration averaged across the flow from the fine-grid solution is subtracted from the coarse Todd and Longstaff prediction as a function of *x* (distance along the flow) divided by *t* (time). In the example discussed in the text, we compute the mean error and covariance matrix at viscosity ratios 5, 10, and 15, and interpolation is used to predict the behavior in between these values.

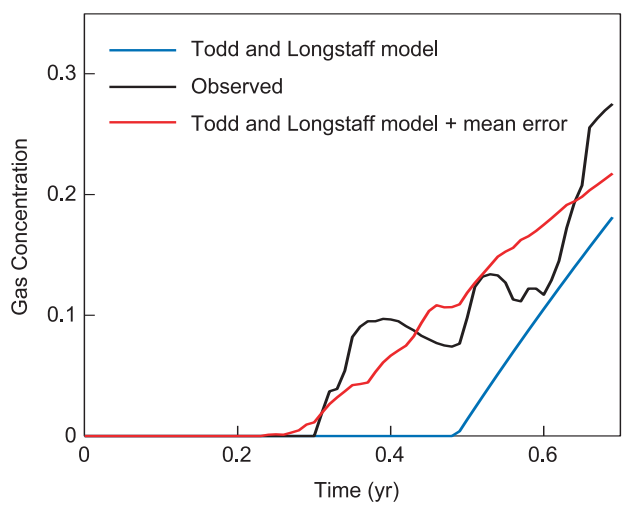


Figure 9. Observed Production Compared with Predictions
The mean error from the error model is added to the coarse-grid result (blue curve) at each time to generate an improved estimate of the gas concentration produced (red curve). The black curve is observed data (actually synthetic data calculated using the fine-grid model with oil-gas viscosity ratio equal to 13).

observed production data (black curve) for which we wish to determine the unknown viscosity ratio. We first run the Todd and Longstaff model at different viscosity ratios from a prior range of 5 to 25 and then correct each prediction by adding to it the mean error at that specific viscosity ratio. The mean error at each viscosity ratio is calculated by interpolating between the mean error at the known calibration points for each value of the similarity variable *x/t*. The blue curve in Figure 9 gives an example of a Todd and Longstaff prediction, and the red curve gives the corrected curve obtained by adding the mean error to the blue curve. To apply the error model, we have converted from the similarity variable *x/t* to time using the known length of the system.

After calculating the corrected predictions for each viscosity ratio, the next step is to compare the corrected prediction (an example is shown in red) for each viscosity ratio with the observed data (shown in black) and compute the misfit M between the simulation and the data. The misfit is given by

$$M = \frac{1}{2}\left(o - s + \bar{e}\right)^T C^{-1}\left(o - s + \bar{e}\right) , \quad (3)$$

where o is the observed value, s is the simulated value, $\bar{e}$ is the mean error, and the covariance matrix is given by $C = \sigma_d^2 I + C_{sem}$. That is, for the covariance matrix, we assume that the data errors are Gaussian, independent, and identically distributed and that therefore they have a standard deviation of σ_d^2, and we estimate the solution error model covariance matrix C_{sem} from the fine-scale simulations performed at the calibration points. The red curve in Figure 10 shows the misfits as a function of viscosity ratio computed using the full error model as in Equation (3). The other misfit statis-

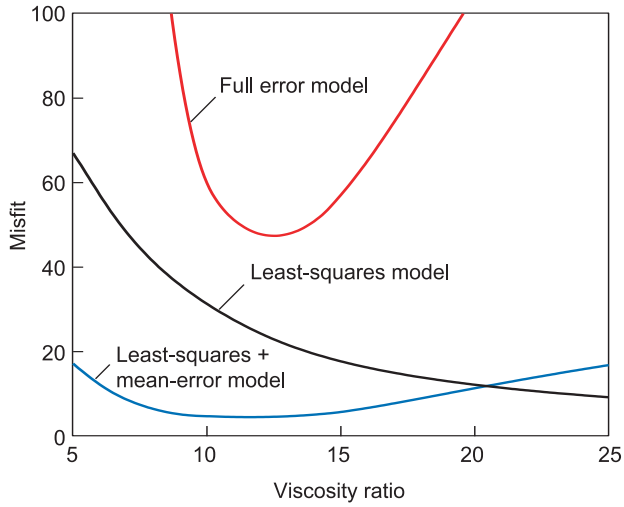


Figure 10. Misfit Statistic vs Viscosity Ratio Calculated in Three Ways
This figure shows a plot of misfit as a function of the viscosity ratio. The misfit is computed using a standard least-squares approach (black curve), least squares with mean error added (blue curve), and the full error model. The misfit measures the quality of the fit to the observed data with low misfits indicating a good fit.

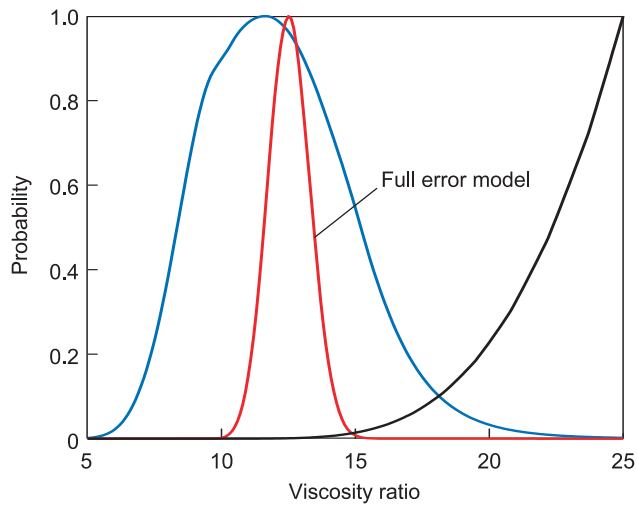


Figure 11. Posterior Probability Distribution Functions for the Viscosity Ratio Calculated in Three Ways
This figure shows the estimated posterior probability (assuming a uniform prior probability in the range 5–25) of the viscosity ratio obtained from three different methods for matching the Todd and Longstaff predictions to observed data. The black curve is obtained from the Todd and Longstaff predictions and a standard least-squares approach. The probability density rises to a maximum at the upper end of the viscosity range specified in the prior model. The blue curve shows the effect of adding the mean error to the predictions. The bias in the coarse model has been removed, but the uncertainty is still large. The red curve shows the estimated viscosity ratio from a full error model treatment—refer to Equation (3)—indicating that it is possible to use a statistical model of solution error to get a good estimate of a physical parameter. The true value of the viscosity ratio in this example was 13.

tics in Figure 10 were computed using

$$M = \sum (o-s)^2 / \sigma_{\mathrm{d}}^2$$

for the least-squares model and

$$M = \sum (o-s+\bar{e})^2 / \sigma_{\mathrm{d}}^2$$

for least-squares plus mean-error model.

The likelihood function L for the viscosity ratio is then given by $L = \exp(-M)$. Notice that the exponential is a signal that the probabilities are sensitive to the method used for computing the misfit. The likelihoods are converted to probability distribution functions by being normalized so that they integrate to 1.

To illustrate the improvement in parameter estimation that results from using an error model, we computed estimates of the probability distribution function for the unknown viscosity ratio using the three different misfit curves in Figure 10, which were calculated with the three different methods: standard least squares, least squares modified by the addition of a mean error term, and least squares with the inclusion of the mean error plus the full covariance matrix. The range of possible values for the viscosity ratio and their posterior probabilities are shown in Figure 11.

The true value of the viscosity ratio used to generate the "observed" (synthetic) production data in Figure 9 was 13, and one can see that this value has been accurately identified by the full error model. The standard least-squares method has not identified this value because of the underlying bias in the Todd and Longstaff model.

We sample from the estimated probability distribution for the viscosity ratio to generate a forecast of uncertainty in future production. Figure 12 is a plot of the maximum likelihood prediction from the Todd and Longstaff model, along with the

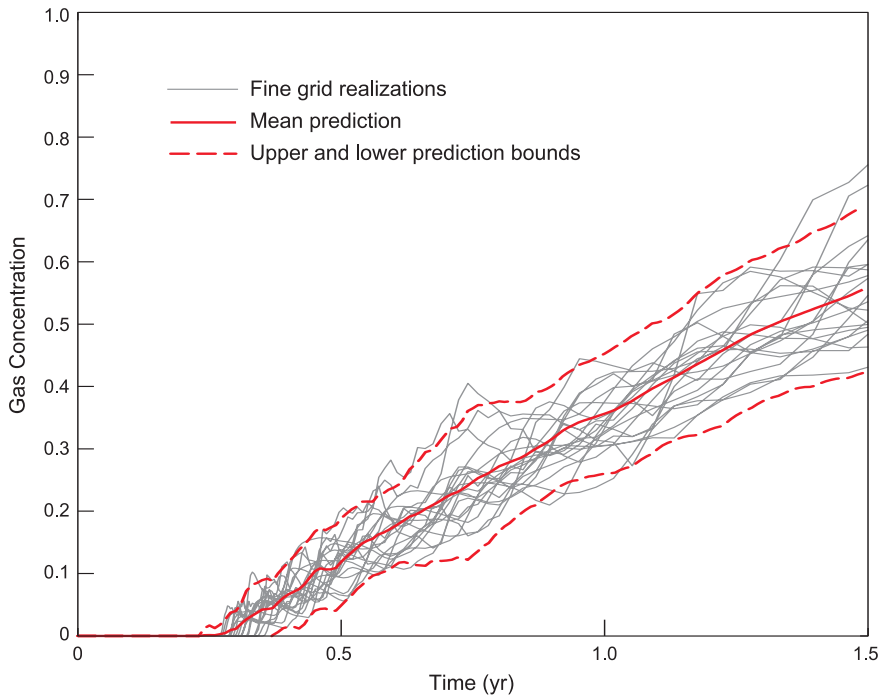


Figure 12. Prediction of Future Oil Production Using Error Model
The solid red line shows the mean (maximum likelihood) prediction from the Todd and Longstaff model and the full error model. The dashed red lines show the 95% confidence interval, and the fine gray curves show the results from 20 fine-grid simulations using the exact viscosity ratio of 13.

95 percent confidence limits obtained by sampling for different values of viscosity. In addition, 20 predictions from fine-grid simulation are shown. They use the exact viscosity ratio 13. The uncertainty in the evolution of the fingers gives rise to the uncertainty in prediction shown by the multiple light-gray curves. It is clear from the figure that use of an error model has allowed us to produce well-calibrated predictions.

Fluid Dynamics—Error Models for Reverberating Shock Waves. Compressible flow exhibits remarkable phenomena, one of the most striking being shock waves, which are propagating disturbances characterized by sudden and often large jumps in the flow variables across the wave front (Courant and Friedrichs 1967). In fact, for inviscid flows, these jumps are represented as mathematical discontinuities. Shock waves play a prominent role in explosions, supersonic aerodynamics, inertial confinement fusion, and numerous other problems. Most problems of practical importance involve two- or three-dimensional (2-D or 3-D) flows, complex wave interactions, and other complications, so that a quantitative description of the flow can be obtained only by solving the fluid-flow equations numerically. The ability to numerically simulate complex flows is a triumph of modern science, but such simulations, like all numerical solutions, are only approximate. The errors in the numerical solution can be significant, especially when the computations use moderate to coarse computational grids as is often necessary for real-world problems. In this section, we sketch an approach to estimating these errors.

Our approach makes heavy use of the fact that shock waves are persistent, highly localized wave disturbances. In this case, "persistent" means that shock waves propagate as locally steady-state wave fronts that can be modified only by interactions with other waves or unsteady flows. Generally, interactions consist of collisions with other shock waves, boundaries, or material interfaces. The phrase "highly localized" refers to shock fronts being sharp and their interactions occurring in limited regions of space and time and possibly being characterized by the refraction of shock fronts into multiple wave fronts of different families. These properties are illustrated in Figure 13, which shows a sequence of wave interactions being initiated when a shock incident from the left collides with a contact located a short distance from a reflecting wall at the right boundary in the figure. Each collision event produces three outgoing waves: a transmitted shock, a contact discontinuity, and a reflected shock or rarefaction wave. The buildup of a complex space-time pattern due to the multiple wave interactions is evident.

Generally, solution errors are determined by comparison to a fiducial solution, that is, a solution that is accepted, not necessarily as perfect, but as "correct enough" for the problem being studied. But producing a fiducial solution may not be easy. In principle, one might obtain one using a very highly resolved computation. However, in real-world problems, this is generally not feasible. If it were, one would just do it and forget about solution errors. So, what do we do when we cannot compute a fiducial solution?

The development of models for error generation and propagation offers an approach for dealing with flows that are too complex for direct computation of a fiducial solution. For compressible flows, the key point is that the equations are hyperbolic, which implies that errors are largely

advected through smooth-flow regions and significant errors are only created when wave fronts collide. The flow shown in Figure 13 consists of a sequence of binary wave interactions, each of which is simple enough to be computed on an ultrafine grid. The basic idea is to determine the solution errors for an elementary wave interaction and to construct "composition laws" that give the error at any given point in terms of the error generated at each of the elementary wave interactions in its domain of influence.

A number of points need to be made here. First, there are a limited number of types of elementary wave interactions. One-dimensional (1-D) interactions occur as refractions of pairs of parallel wave fronts, 2-D interactions are refractions of two oblique wave fronts, and 3-D interactions correspond to triple points produced by three interacting waves. It is important to note that, in each spatial dimension, the elementary wave interactions occur at isolated points. Most of the types of wave interactions that can occur in 1-D flow appear in Figure 13. The coherent traveling wave interactions that occur in 2-D flows have been characterized (Glimm et al. 1985). However, substantial limitations are left on the refinement and thoroughness with which 3-D elementary wave interactions can be studied.

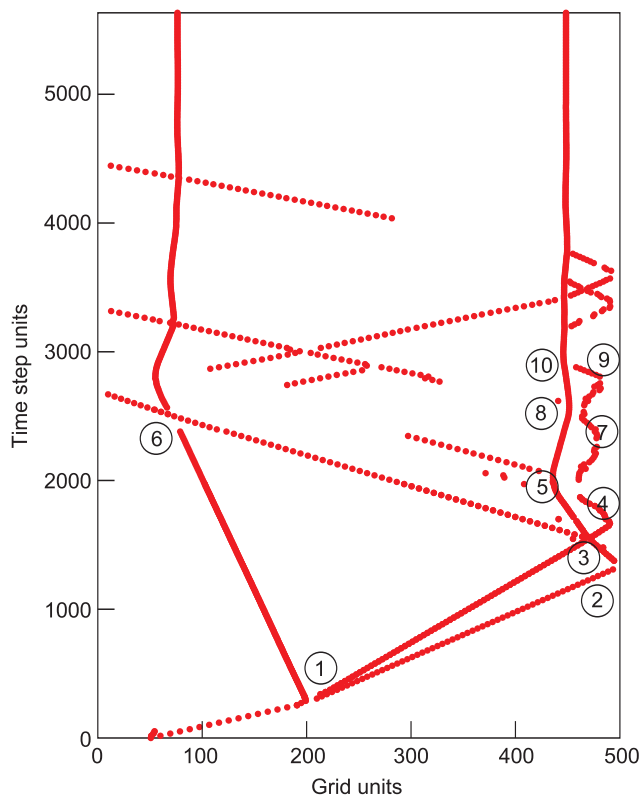


Figure 13. The Space-Time Interaction History of a Shock-Tube Refraction

This figure shows the interaction history as reconstructed from the simulated solution data from a shock-tube refraction problem. A planar shock is incident from the left on a contact discontinuity located near the middle of the test section of the shock tube. A reflecting wall is located on the right side of the tube. Event 1 corresponds to the initial refraction of the shock wave into reflected and transmitted waves, event 2 occurs when the transmitted shock produced by interaction 1 reflects at the right wall, and the events numbered 3–10 correspond to subsequent wave interactions between the various waves produced by earlier refractions or reflections. Our error model is applied at each interaction location to estimate the additional solution error produced by the interaction.

(This figure was supplied courtesy of Dr. Yan Yu, Stony Brook University.)

Event 1 in Figure 13 is a typical example of a 1-D wave interaction. Here, the "incoming waves" consist of an incident shock and a contact discontinuity, and the "outgoing state" is described by a reflected shock, a (moving) contact, and a transmitted shock. The interaction can be described as the solution to a Riemann problem with data given by the states behind the incoming wave fronts. A Riemann problem is defined as the initial value problem for a hyperbolic system of conservation laws with scale-invariant initial data. Riemann problems and their solutions are basic theoretical tools in the study of shock dynamics, the development of shock-capturing schemes to numerically compute flows, and they also play a key role in our study of solution errors. A key point in the use of Riemann problem solutions in our error model is that the solution of a 1-D Riemann problem for hydrodynamics reduces to solving a single, relatively simple algebraic equation. It is thus possible to solve large numbers of Riemann problems for a flow analysis quickly and efficiently. This observation is particularly important because our error model requires the solution of multiple Riemann problems whose data are drawn from statistical ensembles of initial data to represent uncertainties in the incoming waves.

A final point here is that a realistic solution error model must include the study of the size distribution of errors over an ensemble of problems, in which the variability of problem char-

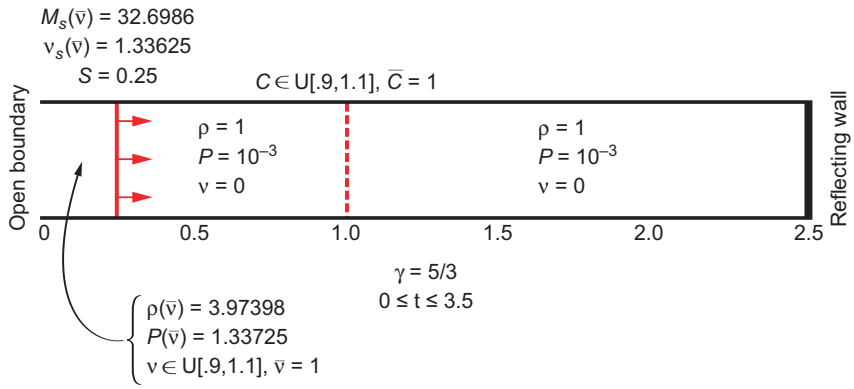


Figure 14. Initial Data for a 1-D Shock-Tube Refraction Problem
This schematic diagram is for the initial data used to conduct an ensemble of simulations of a 1-D shock tube refraction. Each simulation consisted of a shock wave incident from the left on a contact discontinuity between gases at the indicated pressures and densities. Each realization from the ensemble is obtained by selecting a shock strength consistent with a velocity *v* behind the incident shock taken from a 10% uniform distribution about the mean value $\bar{v}$ = 1, and an initial contact location *C* chosen from a 10% uniform distribution about the mean position $\bar{C}$ = 1. In the diagram, *S* is the shock position, M_s is the shock strength, and v_s is the velocity of the shock. The initial state behind the shock is set by using the Rankine-Hugoniot conditions for the realization shock strength and the specified state ahead of the shock.

acteristics is described probabilistically. Of course, one will often want to make as refined an error analysis as possible within a given realization from the ensemble (that is, a deterministic error analysis), but there are powerful reasons for a probabilistic analysis to be needed as well. First, you need probability to describe features of a problem that are too complex for feasible deterministic analysis. Thus, fine details of error generation in complex flows are modeled as random, just as are some details of the operation of measuring instruments. Second, a sensitivity analysis is needed to determine the robustness of the conclusions of a deterministic error analysis to parameter variation. To get an accurate picture, one needs to do sensitivity analysis probabilistically, to answer the question of how likely the parameter variations are that lead to computed changes in the errors. Third, to be a useful tool, the error model must be applicable to a reasonable range of conditions and problems. The only way we are aware of for achieving these goals is to base the error model on a study of an ensemble of problems that reflects the degree of variability one expects to encounter in practice. Of course, the choice of such an ensemble reflects scientific judgment and is an ongoing part of our effort.

Now, let us return to the analysis of solution errors in elementary wave interactions. Our work was motivated by a study of a shock-contact interaction—refer to event 1 in Figure 13. The basic setup is shown in Figure 14, which illustrates a classic shock-tube experiment. An ensemble of problems was generated by sampling from uniform probability distributions (±10 percent about nominal values) for the initial shock strength and the contact position. The solution errors were analyzed by computing the difference between coarse to moderate grid solutions and a very fine grid solution (1000 cells). Error statistics are shown in Figure 15 for a 100-cell grid (moderate grid) solution. Two facts about these solution errors are apparent. First, the solution errors follow the same pattern as the solution (the shock waves) itself; they are concentrated along the wave fronts, where steep gradients in the solution occur. Second, errors are generated at the location of wave interactions. The error generated by the interaction increments the error in the outgoing waves, which is inherited from errors in the incoming waves.

Comparable studies have been carried out for each of the types of wave interaction shown in Figure 13, as well as corresponding wave interactions that occur in spherical implosions or explosions (Dutta et al. 2004). An analysis of statistical ensembles of such interactions has led us to suggest the following scheme for estimating the solution errors. The key steps are (a) identification of the main wave fronts in a flow, (b) determination of the times and locations of wave interactions, and (c) approximate evaluation of the errors generated during the interactions. Wave fronts are most simply identified as regions of large flow gradients, and the distribution of the wave positions and velocities are found by solving Riemann problems whose data are taken from ensembles of state information near the detected wave fronts. The error generated during an interaction is fit by a linear expression in the uncertainties of the incoming wave's strength. The coefficients are computed using a least-squares fit to the distribution of outgoing wave strengths. This fitting procedure can be thought of as defining an input/output relation between errors in incoming and outgoing waves.

A linear relation of this kind, which amounts to treating the errors perturbatively, holds even for strong, and hence nonlinear, wave interactions. But there are limitations. Linearity works if the errors in the incoming waves are not too large, but it may break down for larger errors. In the latter case, higher order (for example, bilinear or rational) terms in the expansion may be needed. See

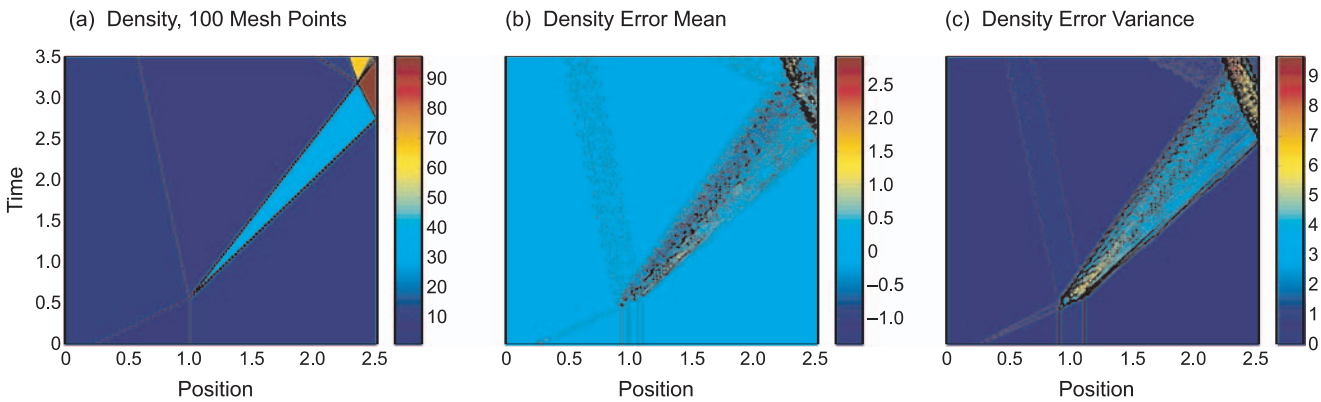


Figure 15. Space-Time Error Statistics for Shock-Tube Refraction Problems
Panel (a) shows the space-time 100-mesh-point density field for a single realization from the flow ensemble. The space-time error field for each realization is computed from the difference between a 100-mesh-zone calculation and a fiducial solution computed using 1000 mesh zones. Panels (b) and (c) show the mean and variance, respectively, over the ensemble as a function of space and time. Note that most errors are generated at the wave interactions and then move with the wave fronts.

Glimm et al. (2003) for details.

We can now explain how the composition law for solution errors actually works. The basic idea is that errors are introduced into the problem by two mechanisms: input errors that are present in waves that initiate the sequence of wave interactions—see the incoming waves for event 1 in Figure 13—and errors generated at each interaction site. However they are introduced, errors advect with the flow and are transferred at each interaction site by computable relations.

Generally, waves arrive at a given space-time point by more than one path. Referring again to Figure 13, suppose you want to find the errors in the output waves for event 3, where the shock reflected off the wall reshocks the contact. On path A, the error propagates directly from the output of interaction 1 along the path of the contact, where it forms part of the input error for event 3. On path B, the output error in the transmitted shock from event 1 follows the transmitted shock to the wall, where it is reflected and then re-crosses the contact. In this way, the error coming into event 3 is given as a sum of terms, with each term labeled by a sequence of wave interactions and of waves connecting these interactions. Moreover, each term can be computed on the basis of elementary wave interactions and does not require the full solution of the numerical problem. The final step in the process is to compute the errors in the output waves at event 3, by using the input/output relations developed for this type of wave interaction.

This procedure represents a substantial reduction in the difficulty of the error analysis problem, and we must ask whether it actually works. Full validation requires use in practice, of course. As a first validation step, we compute the error in two ways. First, we compute the error directly by comparing very fine and coarse-grid simulations for an entire wave pattern. Results are shown in Figure 15. Second, we compute the error using the composition law procedure shown in Figure 13. Comparing the errors computed in these two ways provides the basis for validation.

In Glimm et al. (2003) and Dutta et al. (2004), we carried out such validation studies for planar and spherical shock-wave reverberation problems. As an example, for events 1 to 3 in the planar problem in Figure 13, we considered three grid levels, the finest (5000 cells) defining the fiducial solution, and the other two representing "resolved" (500 cells) and "under-resolved" (100 cells) solutions for this problem. We introduced a 10 percent initial input uncertainty to define the ensemble of problems to be examined. The results can be summarized briefly as follows. For the resolved case, the composition law gave accurate results for the errors (as determined by direct fine-to-coarse grid comparisons) in all cases: wave strength, wave width, and wave position errors. This was not the case for the under-resolved simulation. Although the composition law gave good results for wave strength and wave width errors, it gave poor results for wave position errors. The nature of these results can be understood in terms of a breakdown in some of the modeling assumptions used in the analysis.

An interesting point of contrast emerged between the planar and spherical cases. For the planar case, the dominant source of error was from initial uncertainty, while for the spherical symmetry case, the dominant source of error arose in the simulation itself, and especially from shock

reflections off the center of symmetry.

We come now to the "so what?" question for error models. What are they good for? Our analysis shows that, with an error model, one can determine the relative importance of input and solution errors (thereby allocating resources effectively to their reduction), as well as the precise source of the solution error (for the same purpose), and, finally, one can assess the error in a far more efficient manner than by direct comparison with a highly refined computation of the full problem.

A significant limitation in our results to date is that they pertain mostly to 1-D flows, namely, to flows having planar, cylindrical, or spherical symmetry. Two-dimensional problems are currently under study, while full 3-D problems are to be solved in the future. Furthermore, errors in some important fluid flows lie outside the framework we have developed, and their analysis will require new ideas. One such problem—fluid mixing—was discussed in the previous subsection.

Conclusions

This paper started from the premise that predictive simulations of complex phenomena will increasingly be called upon to support high-consequence decisions, for which confidence in the answer is essential. Many factors limit the accuracy of simulations of complex phenomena, one of the most important being the sparsity of relevant, high-quality data. Other factors include incomplete or insufficiently accurate models, inaccurate solutions of the governing equations in the model, and the need to integrate the diverse and numerous components of a complex simulation into a coherent whole. Error analysis by itself does not circumvent these limitations. It is a way to estimate the level of confidence that can be placed in a simulation-based prediction on the basis of a careful analysis of the source and size of errors affecting this prediction. Thus, the metric of success of an error analysis is the confidence it gives that the errors are of a specific size—not necessarily that they are small (they might not be).

We have reviewed some of the ideas and methods that are used in the study of simulation errors and have presented three examples illustrating how these methods can be used. The examples show how an improved physics model can dramatically reduce the size of errors, how an improved error model can reduce uncertainty in prediction of future oil production, and how an error model for a complex shock-wave problem can be built up from an error analysis of its components.

Similar to models of natural phenomena, error models will never be perfect. Estimates of errors and uncertainties are always provisional because the data supporting these estimates are derived from a limited range of experience. Certainty is not in the picture. Nevertheless, confidence in predictions can be derived from the scope and power of the theory and solution methods that are being used. Scope refers to the number and variety of cases in which a theory has been tested. Scope is important in building confidence that one has identified the factors limiting the applicability of the theory. Power is judged by comparing what is put into the simulation with what comes out.

Error models contribute to confidence by clarifying what we do and do not understand. They also guide efforts to improve our understanding by focusing on factors that are the leading sources of error. Thus, in predictions of complex phenomena, an error analysis will form an indispensable part of the answer. ■

Further Reading

Courant, R., and K. O. Friedrichs. 1967. *Supersonic Flow and Shock Waves*. New York: Springer-Verlag.

Dutta, S., E. George, J. Glimm, J. W. Grove, H. Jin, T. Lee, et al. 2004. Shock Wave Interactions in Spherical and Perturbed Spherical Geometries, Los Alamos National Laboratory document LA-UR-04-2989. (Submitted to *Nonlinear Anal*.).

Gaver, D. P. 1992. *Combining Information: Statistical Issues and Opportunities for Research* Report by Panel on Statistical Issues and Opportunities for Research in the Combination of Information, Committee on Applied and Theoretical Statistics, National Research Council. Washington, DC: National Academic Press.

Gilovich, T., D. Griffin, and D. Kahneman, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge, UK: Cambridge University Press.

Glimm, J., J. W. Grove, Y. Kang, T. W. Lee, X. Li, D. H. Sharp, et al. *2003.* Statistical Riemann Problems and a Composition Law for Errors in Numerical Solutions of Shock Physics Problems, Los Alamos National Laboratory document LA-UR-03-2921. *SIAM J. Sci. Comput.* (in press).

Glimm, J., C. Klingenberg, O. McBryan, B. Plohr, S. Yaniv, and D. H. Sharp. 1985. Front Tracking and Two-Dimensional Riemann Problems. *Adv. Appl. Math.* **6**: 259.

Johnson, V. E., T. L. Graves, M. S. Hamada, and C. S. Reese. 2003. A Hierarchical Model for Estimating the Reliability of Complex Systems. In *Bayesian Statistics 7*: *Proceedings of the Seventh Valencia International Meeting*. p. 199. UK: Oxford University Press.

Kao, J., D. Flicker, R. Henninger, S. Frey, M. Ghil, and K. Ide. 2004. Data Assimilation with an extended Kalman Filter for Impact-Produced Shock-Wave Dynamics. *J. Comp. Phys.* **196** (2): 705.

O'Nions, K., R. Pitman, and C. Marsh. 2002. The Science of Nuclear Warheads. *Nature* **415**: 853.

O'Sullivan, A. E. Modelling Simulation Error for Improved Reservoir Prediction Ph.D. thesis, Heriot-Watt University, Edinburgh.

Palmer, T. N. 2000. Predicting Uncertainty in Forecasts of Weather and Climate. *Rep. Prog. Phys.* **63**: 71.

Sharp, D. H., and M. Wood-Schulz. 2003. QMU and Nuclear Weapons Certification. *Los Alamos Science* **28**: 47.

For further information, contact David H. Sharp (505) 667-5266 (dcso@lanl.gov).

Reducing Uncertainty in Nuclear Data

Mark B. Chadwick, Patrick Talou, and Toshihiko Kawano

Good detective work, combined with theory, experiments, and Bayesian analysis, has reduced by an order of magnitude the uncertainties in the evaluated rate of neutron-induced fission. That reduction allows more accurate simulation of weapon performance. Similarly, more accurate determination of neutron reactions on radiochemical neutron detectors has increased the capability to evaluate the results of past nuclear tests. In both instances, integral experiments with the critical assembly Jezebel are playing an invaluable role. Jezebel and Godiva are the infamous "unclad ladies" from the 1950s. Pictured at left, Jezebel consists of three components of a plutonium sphere that, when brought together, form a critical mass. Unclad, or not encased in neutron reflectors, Jezebel still can support a fast chain reaction with a hard neutron spectrum, characteristic of various nuclear devices.

Weapons performance depends directly on the rates of nuclear reactions, among which the neutron-induced fission chain reaction, shown schematically in the background on the opposite page, is one of the most important. The rates of neutron-induced fission and other neutron-induced nuclear reactions have been measured in numerous experiments. In this article, we describe a project to assess and reduce uncertainties in those basic reaction rates and thereby increase confidence in the predictions of Los Alamos weapons simulation codes (see the box "Uncertainty Quantification for Weapons Certification"). The rate of a nuclear reaction, or more precisely, the cross section for an incident particle to collide and interact with a nucleus[1], varies with the energy of the incident particle. For that reason, cross sections are typically measured at specific incident energies, and the measured values serve as input to the simulation codes. Any uncertainties in those energy-specific, or differential, cross sections translate into uncertainties in the prediction of the overall yield (total energy released) of a nuclear device and other "integral" quantities, so called because they result from the sum of repeated occurrences of the nuclear reaction over a range of incident energies and, in some cases, over the volume of the nuclear material. We present work on reducing uncertainties in two cross sections, both describing neutron-induced processes that are significant for weapon certification: the plutonium fission cross section (see Figure 1), which determines neutron multiplication in a plutonium fission chain reaction, and the cross section for iridium-193 to become the isomer iridium-193m (a long-lived excited state) through neutron inelastic scattering.[2] That process ($^{193}\mathrm{Ir} + n \rightarrow {}^{193\mathrm{m}}\mathrm{Ir} + n'$) has played an important role in diagnosing weapons performance in past underground nuclear tests.

Our work on reducing fission data uncertainties for weapon certification is having an impact on other nuclear technologies. The GEN-IV nuclear reactor program is one such example. This program is exploring several future reactor concepts: more complete burnup of nuclear fuel, proliferation-resistant fuel cycles, and using the reactor as a "waste burner" to transmute long-lived radioactive nuclei into short-lived ones. When the long-time behavior of a GEN-IV reactor was simulated taking into account the best nuclear data available, the known uncertainties in the fission rates led to significant uncertainties in some of the key performance quantities such as nuclear criticality and transmutation rates. Both depend heavily on the fission rates for uranium, plutonium, and several minor actinides (neptunium, americium, and curium). On the basis of this finding, the Advanced Fuel Cycle Initiative Program at Los Alamos is supporting experimental and theoretical research to improve the highest-priority nuclear cross sections—particularly those of the minor actinides that are not currently understood.

Fission cross sections also matter to the nuclear-powered space mission to study Jupiter's moons. With the Laboratory's help, NASA is designing a compact nuclear reactor that will use highly enriched uranium (HEU) to power the plasma thrust engine. The energy output, criticality, radiation environment, and other important features of this reactor are predicted with radiation transport codes that simulate the production of neutrons by the fission process and their subsequent movement and participation in fission and other nuclear interactions. Even though the mission to Jupiter would be unmanned, a safe launch is most important; at the same time, we must also be able to guarantee that, if a crash were to occur, the probability of a criticality accident would be negligible. This project, therefore, also needs estimates of fission cross-section uncertainties—in this case, uranium-235 fission—to guide design of the space reactor.

Both statistical analyses of data

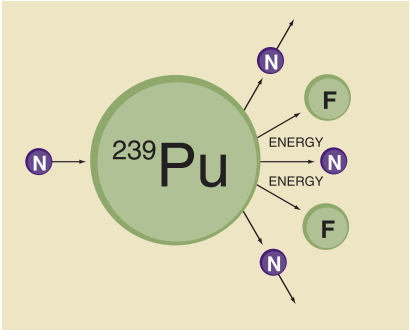


Figure 1. Schematic of Neutron-Induced Fission of Plutonium-239
This artist's conception of the fission process (as well as the simplified fission chain reaction in the background on the opposite page) shows an incoming neutron (purple) being absorbed by a plutonium-239 nucleus, which causes the nucleus to split into two 'fission' fragments (small green circles) and release several neutrons (purple). In reality, the nucleus first splits into two highly excited fragments and then each fragment releases one or more neutrons. The resulting fission fragments are typically radioactive nuclei and sometimes release additional ('delayed') neutrons. Both the 'prompt' and delayed neutrons can induce fission in nearby plutonium-239 nuclei, causing a fission chain reaction.

[1] A nuclear collision cross section $\sigma(E)$ measures the probability for an incident particle of energy E, say a neutron n, to collide and interact with or scatter from a nucleus N and produce some final state.

[2] In inelastic neutron scattering, the incident neutron n transfers energy to the nucleus N and leaves with less energy. That process ($N + n \rightarrow N + n'$) is denoted (n,n'), where the left neutron is incoming, the right neutron is outgoing, and the prime indicates that the outgoing neutron has a different energy than the incoming one.

Uncertainty Quantification for Weapons Certification

Since the end of nuclear testing, the Department of Energy has focused on developing a set of weapons simulation codes that more accurately model weapon explosions. This Advanced Simulation and Computing (ASC) Program has several objectives: creating simulation codes that implement more-accurate algorithms for solving the relevant hydrodynamics and radiation transport equations, building some of the fastest computers in the world on which to run these codes, and developing improved materials and physics models and data for "high-fidelity" weapons simulations. Such new simulation codes are needed to certify the safety and reliability of the U.S. stockpile and to answer questions about aging components in stockpiled weapons.

Quantification of the margins-and-uncertainties (QMU) concept has been adopted as the framework within which certification is performed. At each critical stage in the sequence of a weapon explosion, researchers in the Applied Physics Division at the Laboratory assess margins for certain physical quantities that enable the weapon to perform reliably. The QMU process formalizes the considerations and assumptions that go into modeling a weapon's performance and assessing whether it will perform correctly. A component of QMU is uncertainty quantification, whereby we determine how uncertainties in the underlying physics models and data impact the accuracy of full simulation results for weapons. It is in this context that we are assessing the accuracy of the plutonium fission cross-section data.

from past differential measurements and new state-of-the-art differential measurements at the Los Alamos Neutron Science Center (LANSCE) play a crucial role in allowing us to reduce cross-section uncertainties. More surprising, perhaps, is that small-scale integral experiments performed at the Los Alamos Critical Experiment Facility (LACEF) are having a huge impact in the validation of nuclear data used in weapons codes, as well as in reducing data uncertainties (see Figure 2). In the case of plutonium fission, for example, these criticality experiments have led to a factor of 10 reduction in the predicted fission process uncertainties,[3] as is discussed in more detail below.

As the name implies, a criticality experiment entails very careful assembling of a radioactive target made from special nuclear materials (plutonium, uranium-235 and -238, and other fissile materials) into a critical mass, that is, one that creates a self-sustaining fission chain reaction and a flux of neutrons with energies typical of fission. In fact, the energy spectra of the neutrons within the various assemblies at LACEF have been precisely determined through a combination of theory, simulation with radiation transport codes, and experiment. Thus, despite being integral experiments involving a wide spectrum of neutron energies and very large numbers of fission reactions occurring over a short period, critical-assembly experiments are well-characterized static nuclear physics experiments from which basic cross-section data can be inferred. In contrast, archival data from past Nevada underground nuclear tests were obtained from much more complicated integrated experiments involving hydrodynamics and other phenomena, in addition to nuclear physics.

Over the last few decades, nuclear criticality experiments have been used not only to reduce uncertainties in evaluated nuclear data libraries but also to validate the radiation (neutron and gamma-ray) transport methods used in our particle transport codes for static nuclear devices. One such code is the widely used Monte Carlo *N*-Particle Transport Code (MCNP). Developed by the Diagnostic Methods Group at Los Alamos, MCNP has become the international standard Monte Carlo code for simulating neutron transport and criticality in reactor applications and nuclear criticality safety studies. Nuclear criticality benchmark experiments developed at LACEF produce neutrons with a wide range of energy spectra: Some experiments mimic the highly thermalized systems of standard reactors, producing slow neutrons with an average energy of 0.025 electron volt (or soft neutron spectra); other experiments at the opposite extreme produce fast neutrons with an average energy of 1 to 2 million electron volts (MeV), or hard neutron spectra. The fast critical assemblies at LACEF are particularly relevant for validating our cross-section databases for weapons research because they produce a fast chain reaction (involving energetic neutrons). The Jezebel fast assembly is a critical mass of plutonium with no neutron reflectors, or cladding, the Godiva assembly is another 'unclad' assembly containing a critical mass of HEU, and the Flattop assemblies include cores of plutonium or HEU made critical with reflector materials.

[3] So-called "evaluated" nuclear data result from analyzing all available experiments, resolving discrepancies, and determining both the values and the uncertainties. They are kept in libraries known as ENDF for evaluated nuclear data files.

The two examples discussed below use fast critical-assembly measurements in different ways. In the case of plutonium, it is a precise measurement of the plutonium critical mass that allows us to accurately validate (and reduce the uncertainties on) the plutonium neutron-induced fission cross section, in part because our radiation transport methods in the MCNP code are so accurate. In the case of iridium, samples of iridium are placed at different locations within the critical assembly, and each is irradiated by a different spectrum of neutrons characteristic of its location within the assembly. The neutrons at different locations have not only different distributions of energies but also different mean energies. Thus, measuring iridium reaction rates within different parts of the assembly provides an important validation of the iridium cross sections at different average neutron energies.

Neutron-Induced Fission Cross Section of Plutonium

The neutron-induced fission cross section of plutonium-239 represents the probability that, when a single neutron hits a target nucleus of plutonium-239, the composite system of target plus neutron breaks apart, usually into two smaller nuclei fragments, $n + {}^{239}\mathrm{Pu} \rightarrow$ fission fragments. This probability naturally depends on the kinetic energy of the incident neutron and is therefore represented as a two-dimensional curve of cross section vs neutron energy (see Figure 3). To convert this probability into a rough estimate of the number of plutonium-239 fissions occurring in a real application—for example, in the core of a nuclear reactor—over a given period or in a critical assembly experiment, this cross section averaged over the neutron energies is multiplied by the neutron fluence (the neutron flux integrated over the relevant time).

Figure 2. The Los Alamos Critical Assembly Facility as Seen through an Anasazi Cave

Statistical Analysis of Experimental Data. The theory of nuclear fission has advanced considerably over the last fifty years, and especially within the last decade as high-performance computers made complex calculations feasible (see references by Peter Möller at the end of this article). Nevertheless, theoretical predictions of the neutron-induced fission cross section remain too imprecise for practical calculations of real systems. Experimental measurements of the cross section must therefore be relied on, and the cross section at most incident neutron energies is typically known to about 2 percent accuracy. Until now, however, the fission cross section for incoming neutrons of energies just below 14 MeV was known to only 4 percent accuracy. That deficiency motivated a significant effort to reanalyze the cross-section data from numerous (sometimes discrepant) experiments. We applied statistical methods to evaluate the cross-section data, assessed the resulting uncertainties, and were able to reduce uncertainties considerably.

An experiment typically yields a numerical value of a physical observable, which in turn is related either directly or indirectly to the physical quantity we are interested in. Of course, no experiment is perfect, and information on the uncertainties associated with the measured value is essential for judging the validity of the result. Uncertainties come from multiple sources but are commonly classified into two categories: statistical and systematic. Statistical uncertainties follow the simple $1/\sqrt{N}$ rule; that is, if the same experiment is repeated N times, the statistical uncertainty of the measured value will be proportional to $1/\sqrt{N}$. In the limit of an infinite number of identical experiments, this uncertainty would be null. Such uncertainties reflect inherent fluctuations in the measurement itself, and for a large number of repeated experiments, the measurement fluctuations average to zero.

Systematic uncertainties include all uncertainties other than statistical ones and, unlike the latter type, cannot be indefinitely reduced by repetition of the same experiment. Examples of systematic uncertainties will be given later for the plutonium-239 fission cross section. From the point of view of data analysis, systematic uncertainties define a lower limit for the accuracy of a given experimental setup. This fact alone justifies using different experimental setups to measure the same quantity. Because the sources of systematic errors differ from one experimental setup to another, differences in the results from different setups provide a clue on ways to go beyond the lower limits imposed by each individual experiment. By performing a statistical analysis on data from not only one but several experiments aimed at measuring or inferring the same physical quantity, it is possible to quote a value with an uncertainty smaller than the one of each individual experimental result.

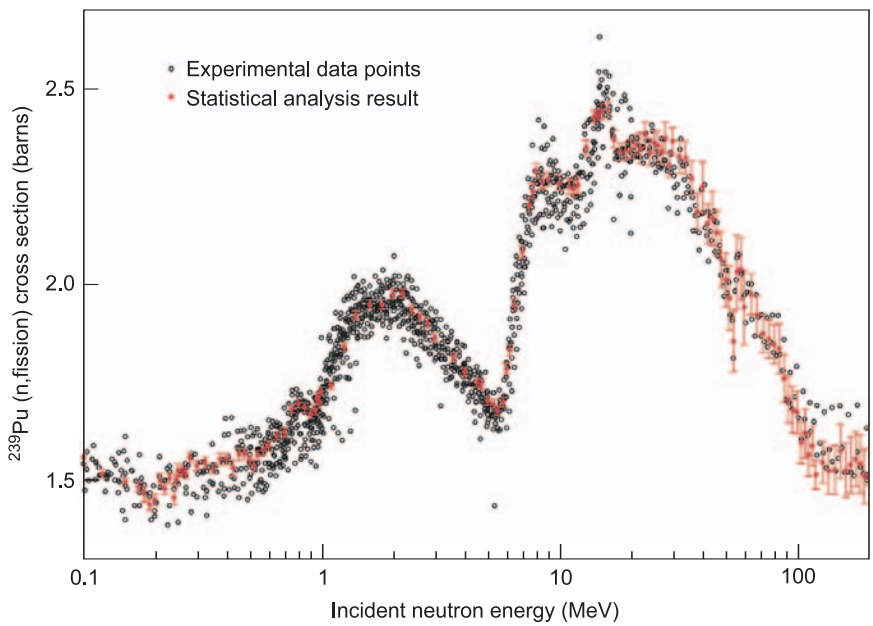


Figure 3. Reducing Uncertainties with Bayesian Statistical Analysis
The neutron-induced fission cross section of plutonium-239 has been measured numerous times over the incident-neutron energy range plotted here. The black dots represent the experimental data from many laboratories (including Los Alamos), which originate either from a direct measurement of the cross section or from a ratio measurement to the well-known neutron-induced fission cross section of uranium-235. (For figure clarity, we did not display the experimental error bars.) The spread of experimental data is a simple indicator of how well the cross section is known. The result of our Bayesian analysis study is shown in red dots, along with the resulting standard deviations. This figure explicitly demonstrates how a Bayesian statistical analysis can help reduce the uncertainties on our knowledge of this important cross section. At higher energies, the error bars tend to increase because two discrepant data sets are present.

In 1763, the work of Reverend Thomas Bayes on inference logic was published posthumously. Based on the theory of conditional probabilities, Bayes' theorem provides a logical and mathematically sound framework to update knowledge in view of new evidence. This concept is paramount in many areas of science and even more generally in any field of study that involves learning algorithms. Simply stated, Bayes' theorem reads

$$\mathcal{P}(\mathcal{H}|\mathcal{D},\mathcal{I}) \propto \mathcal{P}(\mathcal{D}|\mathcal{H}) \times \mathcal{P}(\mathcal{H}|\mathcal{I})$$
$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \quad .$$

The term $\mathcal{P}(\mathcal{H}|\mathcal{I})$, or prior, answers the question, "how probable is the hypothesis $\mathcal{H}$, given the information known prior to the experiment?" In other words, the prior represents the state of our knowledge of (or belief in) the hypothesis $\mathcal{H}$ before the new information, in the form of the data $\mathcal{D}$, is included. The prior is multiplied by $\mathcal{P}(\mathcal{D}|\mathcal{H})$, the likelihood function, which quantifies how important the new data $\mathcal{D}$ are to our overall knowledge of the hypothesis $\mathcal{H}$. The likelihood function answers the question, "how probable is the observation of data $\mathcal{D}$ if the hypothesis $\mathcal{H}$ were actually true?" It provides the central and fundamental link between our prior knowledge and the posterior function $\mathcal{P}(\mathcal{H}|\mathcal{D},\mathcal{I})$, which answers the question, how probable is $\mathcal{H}$, now that we know both $\mathcal{D}$ and $\mathcal{I}$? In other words, the posterior measures the degree of confidence in $\mathcal{H}$ after the new data are taken into account.

Because a Bayesian analysis explicitly contains the concept of a prior knowledge, concerns have been raised about the subjectivity of such an approach, as opposed to more traditional statistical-analysis techniques. A Bayesian analysis is inherently a recursive process, in which information is integrated step by step. This means that the first step relies on a prior that is not based on any real information. When data are scarce, the result of the analysis can be distorted according to the specific choice made for the prior. This type of analysis appears to be in stark contrast with more traditional statistical analyses that are based on only real data. However, the contrast is only apparent, and the supposed flaw in the Bayesian approach seems to be only semantic. In any case, this issue is not relevant to our study: The number of data sets on the neutron-induced fission cross section of plutonium-239 is sufficiently large that the result of our analysis is insensitive to the choice of a particular prior.

The presence of this large data set could also lead us to think that much is known on this particular cross section and that there is no need to investigate further. The truth is not quite that simple. First, it is not uncommon to find discrepant experimental results, that is, results with error bars that do not overlap. Experimental data points for the plutonium-239 fission cross section are shown in Figure 3, illustrating how large the scattering in experimental results can be. Second, information on the uncertainties (and their sources) associated with a given data set is often only partially given, and for some (mostly older) experiments no information is available. As a result, our evaluation is all the more difficult. Finally, whereas most experimental results will be accurate at a

3 to 10 percent level, some important applications that need the plutonium-239 fission cross section require an accuracy closer to 1 to 2 percent. As mentioned earlier, a statistical analysis, Bayesian or otherwise, can help to more precisely determine the fission cross section.

We used a standard Bayesian approach to evaluate the plutonium-239 fission cross section from incident-neutron energies between 0.1 and 150 MeV. This energy range corresponds to a region where the cross section is a fairly smooth function of the incident energy (no resonances) and where the fission channel is dominant compared with other competing neutron-induced processes such as neutron capture, inelastic neutron scattering, and (n,2n) reactions, in which a nucleus absorbs the incoming neutron and promptly emits two.

Although the mathematical toolbox to evaluate the fission cross-section data was in place, inherent in this task was the need to reconstruct the uncertainties and correlations of important unpublished fission measurements that were performed (often many years ago) at numerous facilities around the world. This need required detective work.

In many cases, we were almost completely dependent upon the expertise of senior nuclear-data experimentalists and theorists, many of whom have retired or are close to retirement. These experts have in-depth knowledge of measurements made decades ago and a good (sometimes intuitive!) understanding of which experimentalists and facilities are most reliable.

Sources of experimental uncertainties are numerous and varied, depending on the particular experimental facility, detectors, and measurement and analysis techniques employed. In addition, the measured observable is often some function of the physical quantity of interest rather than the quantity itself. To determine the fission cross-section, for example, one measures the number of fissions produced during neutron irradiation of the target, which is proportional to fission cross section times the neutron fluence (defined as the neutron flux integrated over time). The neutron fluence is quite difficult to measure precisely and therefore introduces a large uncertainty into the results. Consequently, many experiments do not measure the plutonium fission cross section directly. Instead, they measure the ratio of the plutonium-239 to the uranium-235 fission cross section. By measuring that ratio, they eliminate the dependence on the neutron fluence and thus a large source of uncertainty. But a ratio is not a cross section. To come back to the quantity of interest, the experimental result needs to be multiplied by an ‘evaluation’ of the uranium-235 fission cross section, that is, a carefully determined result along with the uncertainties. Uncertainties on this cross section will then act upon the uncertainties on the plutonium-239 fission cross section in a highly correlated manner.

The neutron-induced fission cross section of uranium-235 is denoted as a standard cross section, one that experimentalists can use with confidence to renormalize their results (that is, convert measured ratios into cross sections) because it is a smooth function over a certain energy range and known very accurately. However, our knowledge of this cross section has changed over the years, by up to 2 percent in some energy regions. These differences are large enough that we have had to renormalize, according to current standard values, all the older experimental results obtained with standards known at the time in order to make a direct comparison of experimental data.

There are numerous other sources of uncertainties that must be analyzed. Over the years, different types of detectors have been used to measure or infer neutron-induced fission cross sections: A fission chamber that detects one (or sometimes the two) fission fragment(s), a proton telescope that uses the (n,p) reaction to estimate the neutron fluence, a time-of-flight (TOF) measure of the neutron incident energy, and others. Depending on the particular experimental setup, we have attempted to estimate the uncertainties associated with a given measured result after the fact, even though the experimentalist has recorded only partial or no information regarding error sources. In many cases, experimentalists have reported only statistical errors although a correct estimation of the systematic uncertainties is necessary for obtaining a quality result. Sometimes, we can fill in some of that information. For example, if two experiments have been performed in the same institute, they often use the same neutron source and target samples, in which case we include correlations between the two results in our analysis. In addition, documentation of the experimental details in one case can help us infer the values of uncertainties in the other.

Uncertainties may also exhibit an energy dependence. For example, if a detector efficiency is known to a certain accuracy within a given energy range, that accuracy defines some correlation among the results obtained with that detector within the specific energy range.

All these nonstatistical uncertainties cannot be described by simple standard deviations, but by correlations between fission cross sections at different energies. Correlations in the fission cross section are best represented by the so-called covariance matrix, whose diagonal and off-diagonal elements represent the standard deviations and the correlations, respectively. The off-diagonal elements play a key role in our statistical

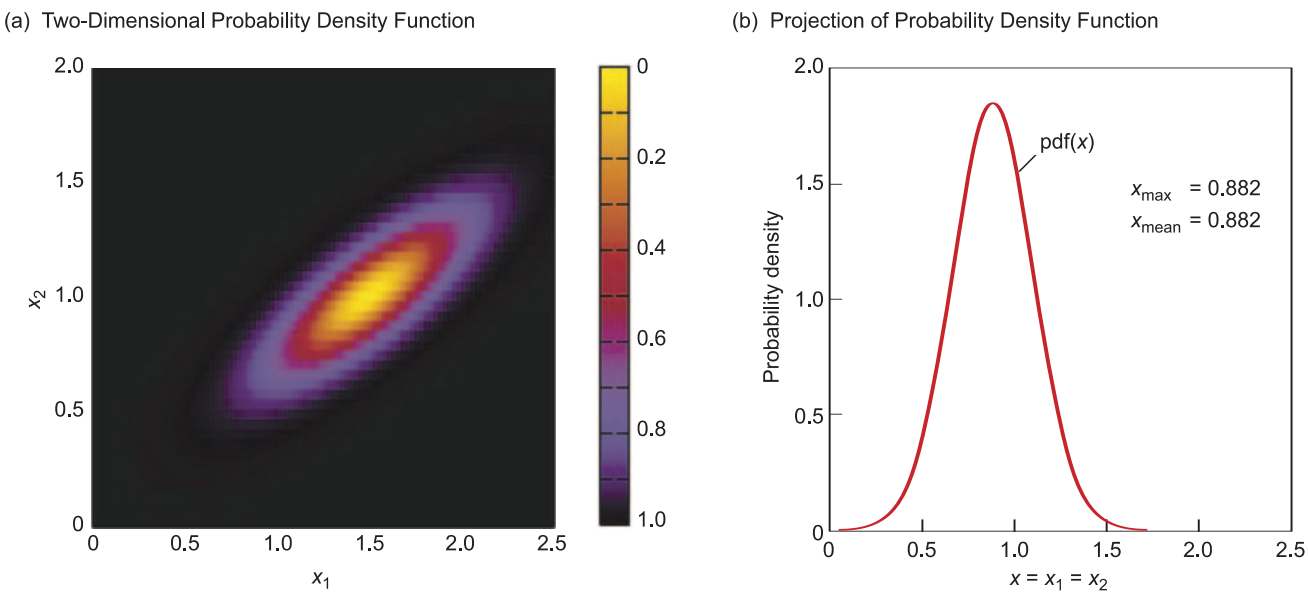


Figure 4. Peelle's Pertinent Puzzle (PPP)

Robert Peelle introduced the puzzle that now bears his name to illustrate the importance of including systematic errors in nuclear data evaluations. In his original example, there are two measurements of the same physical quantity, and the results are 1.5 and 1.0 respectively. Each result has a 10 percent uncertainty, and both results share a 20 percent common error. Standard statistical tools applied to this case give a best-estimate value of 0.882 for the physical quantity, which falls below both measured values! In (a), the two-dimensional Gaussian probability distribution function for the two measurements pdf(x_1,x_2) is shown; in (b), the projection of this distribution is shown on the $x_1 = x_2$ line, with the mean and maximum values equal to 0.882. However, this value depends on an underlying assumption regarding the nature of the correlated uncertainty. In practice, this knowledge is not often available.

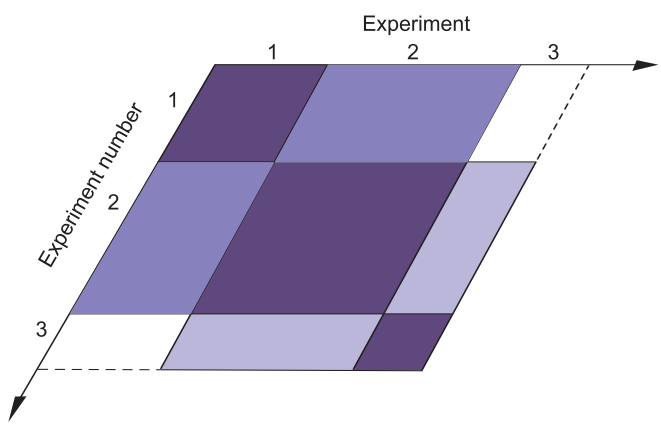


Figure 5. Representation of a Covariance Matrix

The role of a nuclear data evaluator includes constructing covariance matrices that completely describe the experimental data sets and the associated uncertainties and correlations for a given nuclear cross section. Each experimental set corresponds to an ensemble of cross-section values for various incident neutron energies. Statistical uncertainties are commonly given, whereas sources (and quantification) of systematic errors are only sometimes available. Correlations between different data sets can also exist—for example, if the same experiment facility, detector, or sample target is used in two distinct experiments. This picture shows a schematic representation of a corner of the large covariance matrix that results from the study of all the cross section data at all energies.

analysis. Unfortunately, they are also the most challenging quantities to estimate.

A fascinating example of the role of correlations in statistical analysis is Peelle's Pertinent Puzzle, or PPP for short (refer to Figure 4) named after Robert Peelle, who confronted the nuclear data community with a counterintuitive example. Suppose that two measurements of the same physical quantity are made, and the results are 1.5 and 1.0 respectively. Each result has a 10 percent uncertainty, and both results share a 20 percent common error. Standard statistical tools applied to this case give a best-estimate value of 0.882 for the physical quantity, which falls below both measured values! This result may be correct depending on the nature of the correlated uncertainty, additive or multiplicative. Because in many instances

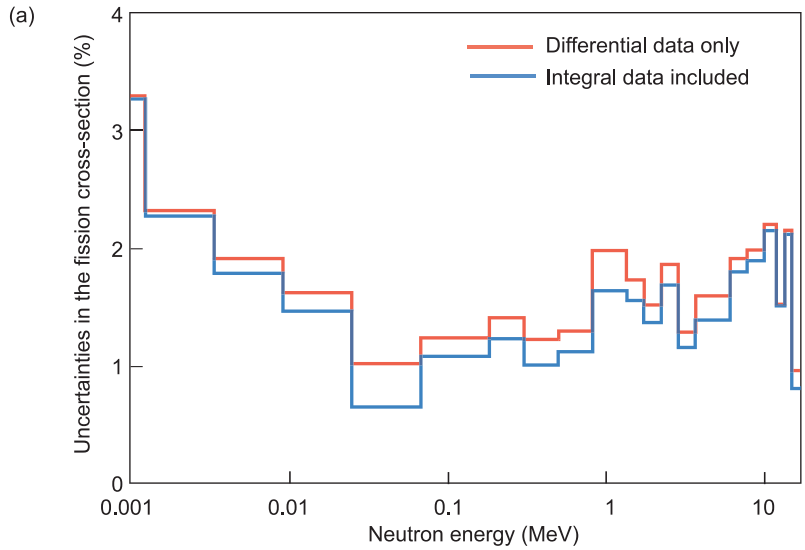
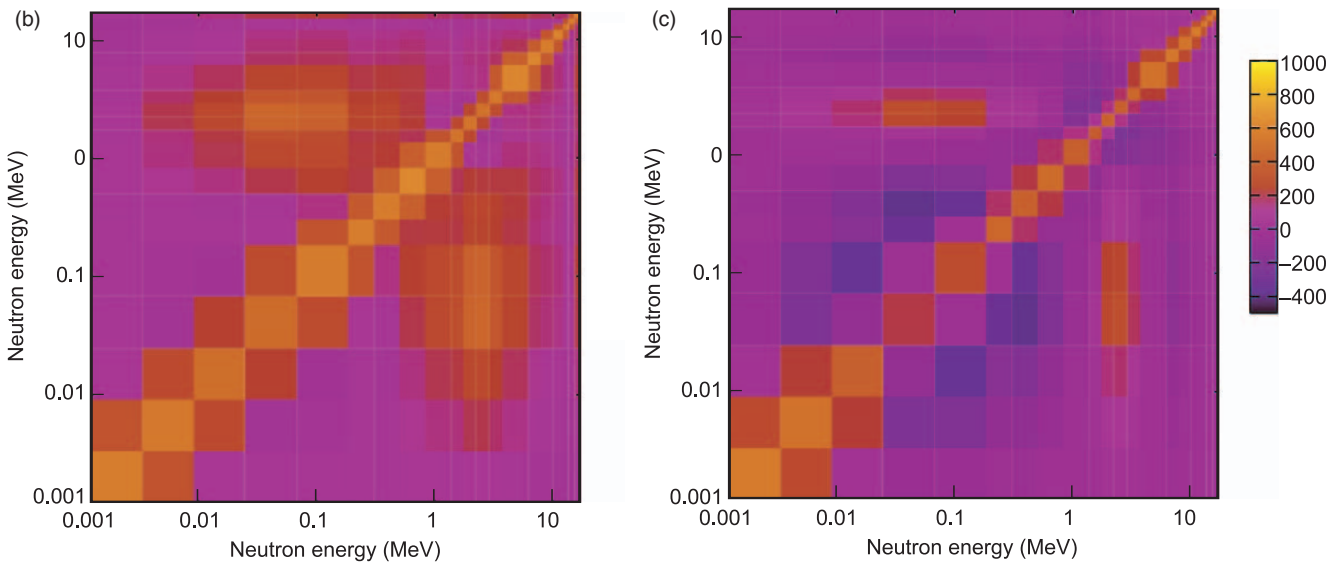


Figure 6. Evaluated Variance-Covariance Matrix for the Pu-239 Fission Cross Section

In (a), the evaluated variances corresponding to the evaluated Pu-239 fission cross section (shown in Figure 3) are given when only differential data are used (red) and when integral data (blue) from critical assembly experiments are also included in the analysis. The corresponding correlation matrices are shown in (b) and (c), before and after inclusion of integral data in the analysis, respectively. The impact of adding integral information into our statistical analysis is clearly seen: It tends to reduce the standard deviations and generate negative correlation values that strongly constrain the fission cross section.



we do not know the origin of uncertainties, PPP represents a real puzzle for nuclear data evaluators, confronted with older and not well-documented experimental data.

The result of our comprehensive statistical analysis is depicted in Figure 3, along with the experimental data sets. The representation of the uncertainty with simple error bars on individual points is only part of the story. The covariance matrix for the evaluated cross section is also quite important. Figure 5 shows a schematic view of a portion of a covariance matrix that represents uncertainties and correlations among all experimental data sets included in the statistical analysis. The actual covariance matrix for the evaluated fission cross section is shown in Figure 6.

In summary, the correct estimation of experimental uncertainties and correlations is undoubtedly the most important aspect of precisely evaluating the plutonium fission cross section (and its uncertainties) in this kind of statistical analysis. Our project has benefited from extensive expertise by scientists at many institutions—especially at Los Alamos, the National Institute of Standards and Technology, and at the International Atomic Energy Agency—that have a long history of understanding and assessing the uncertainties and correlations in previous cross-section measurements.

Critical-Assembly Constraints on Fission Cross-Section Data. We have discussed how uncertainties on the plutonium fission cross sections can be determined from a Bayesian analysis of the experimental cross-section data. Next we show how integral

measurements of the critical mass of plutonium are allowing us to make much larger reductions in uncertainty. Our ability to accurately model a critical assembly of plutonium using the MCNP transport code in conjunction with our neutron cross section data provides constraints on the uncertainties on the underlying microscopic plutonium fission cross-section data.

MCNP was developed at Los Alamos over many decades and is the world's most widely used, sophisticated, and well-tested code for simulating the coupled transport of neutrons and photons as they interact with nuclei. The interactions of neutrons with individual nuclei are modeled using nuclear cross sections from the evaluated neutron data files (ENDF) database developed at Los Alamos and other national laboratories. The accuracy of the transport calculational methods is so high that MCNP simulations of integral experiments, such as the criticality of a sphere of plutonium, provide a valid test of the accuracy of the underlying ENDF nuclear cross sections such as neutron-induced fission.

The calculated critical mass of plutonium depends on cross sections for a number of different neutron-plutonium interactions. It depends on the neutron-induced plutonium fission cross section, the average number of prompt neutrons ($\bar{\nu}$) emitted from fission fragments after a plutonium nucleus fissions, the cross sections for inelastic scattering of neutrons by plutonium nuclei; the angular distributions of neutrons that scatter elastically from a plutonium nucleus; and the cross section for a plutonium nucleus to capture a neutron. Of these quantities, it is the first two, and more precisely the product of the fission cross section and $\bar{\nu}$, that most sensitively influence the calculated critical mass and the neutron multiplication rate k_{eff} in the system, which equals unity when the system is critical.

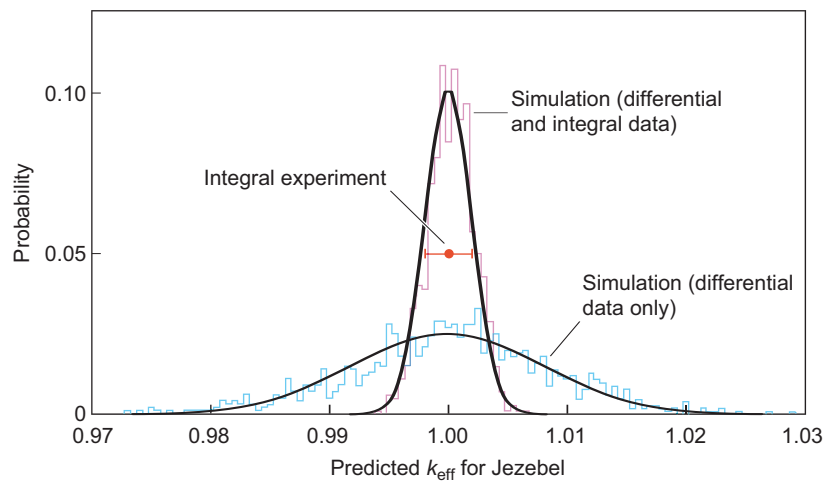


Figure 7. Probability Distribution Function for Jezebel's Neutron Multiplication Rate

We used our Bayesian uncertainty quantification code KALMAN to combine the prior information on differential fissions cross section measurements with the integral information from the Jezebel critical assembly measurement. The analysis provides posterior fission cross sections for different neutron energies. The simulation of Jezebel's criticality using those posterior cross sections yielded a probability distribution for k_{eff} (pink curve) that has a much smaller variance than that of our initial result from differential data only (blue curve).

If we were to estimate the fission cross section and $\bar{\nu}$ uncertainties based on only the fundamental, measured differential cross-section data discussed in the previous section, we would obtain uncertainties in the range of 1 to 2 percent for the fission cross section and less than 1 percent for $\bar{\nu}$, for neutrons with energies in the fission-spectrum energy range of 1 to 2 MeV. In an MCNP transport simulation of Jezebel, these numbers would translate into calculated uncertainties in the range of 1 to 2 percent for calculated values of k_{eff}.

However, Jezebel's measured criticality defines the k_{eff} uncertainty to less than 0.2 percent—an order of magnitude smaller than our previously calculated results based on cross section and $\bar{\nu}$ data uncertainties. We have used those integral measurements which are simple and highly accurate to constrain the differential fission cross sections by using the standard Bayesian technique. With this method, we were able to reduce uncertainties in the fission cross section, and the combined differential and integral data now predict that the neutron multiplication due to fission (k_{eff}) is accurate to about 0.2 percent, an order of magnitude more precise.

The plots in Figure 6 illustrate the uncertainty reductions. The uncertainties (variance and covariance) associated with the statistical analysis of the differential experimental data alone are shown (red line) in Figure 6(a) (the variance) and in Figure 6(b) (the correlation matrix multiplied by 1000). Neutron transport calculations were performed for the Jezebel critical assembly, and the sensitivity coefficients of the cross sections to the neutron multiplicity were obtained. Then Jezebel data were used to adjust the fission cross section through the Bayesian inference method. The resulting uncertainties in the fission cross section are shown in Figure 6(a) (blue line). The impact on the fission cross section itself is very small. However, the uncertainties become

smaller, and negative correlations appear, as shown in Figure 6(c).

These negative correlations constrain the fission cross sections in order to keep the integral quantities constant. If we generate randomly sampled fission cross-section ensembles in accordance with this covariance matrix, the calculated values of k_{eff} for Jezebel form a Gaussian distribution of 0.2 percent uncertainty. This result can be seen in Figure 7, where the large reduction in the uncertainty of the calculated criticality is evident by comparison with the uncertainty from methods that do not use integral measurements.

Iridium Nuclear Cross Sections

Nuclear weapons performance is affected by the neutrons the weapons produce. The neutrons induce nuclear fission in the plutonium and uranium components of the device, and a runaway fission chain reaction occurs that releases the massive amount of energy driving the nuclear explosion. Many of the variables that affect weapons performance depend on the energy distribution (spectrum) of the neutrons. The neutron energy spectrum, for example, determines the relative rate at which fission occurs versus other neutron-induced nuclear reactions, and it also determines the number of neutrons released per atom during the fission process.
Certain elements have been used almost since the inception of the nuclear age to gain spectral information about the all-important neutrons. Small amounts of these so-called radiochemical (radchem) detector materials were placed in specific locations within a nuclear weapon before a test. During the explosion, the intense neutron flux transmuted some of the atoms of the detector material into other, predominantly radioactive, isotopes. After obtaining tiny amounts of the postshot test debris, radiochemists would extract the detector element from the samples and measure the relative amount of each radioactive isotope. Provided the nuclear cross sections for the production and/or destruction of the stable and radioactive isotopes were well understood and measured accurately, a weapons designer could relate the isotopic ratios to the neutron fluence[4] that occurred within the device.

During the era of nuclear weapon testing, different radchem detectors were used to measure the neutron fluence in different energy ranges. Certain nuclear reactions—for example, the $(n,2n)$ reaction, in which one neutron impinges on a nucleus and two neutrons are emitted—are known as threshold reactions; they occur only if the energy of the incident neutron is above some threshold energy, typically a few million electron volts or higher. Isotopes that are produced by the $(n,2n)$ reaction were used to measure the high-energy (about 14 MeV) neutron fluence produced by fusion reactions. Other reactions for producing new isotopes, notably the (n,γ) neutron capture process (in which a nucleus captures an incident neutron and emits a gamma ray), have no threshold. Neutron capture is more likely to occur as the neutron energy decreases and (n,γ) neutron capture reactions dominate isotope production when the neutron energy is below 1 MeV.

The reaction that has been used as a diagnostic for neutron energies between these two extremes is the (n,n') inelastic neutron-scattering reaction in which an iridium-193 nucleus absorbs some energy from the incident neutron and transitions to a long-lived nuclear excited state known as the isomer iridium-193m. This reaction is uniquely sensitive to neutrons with energies in the few-million-electron-volt range, which, in turn, is the energy range of neutrons produced in the fast chain reaction in a weapon. Hence, determining the production of the isomer iridium-193m is an extremely important diagnostic for weapon performance.

Figure 8 indicates with arrows the reaction pathways that can occur when neutrons are incident on an iridium target composed of the stable isotopes 191 and 193. By measuring the production of radioactive iridium-189, -190, -192, 193m, and –194 in such a target, one can learn information about all three energy-sensitive neutron-induced reactions, $(n,2n)$, (n,n'), and (n,γ). Iridium, therefore, provides a unique diagnostic capability of the neutron fluence in multiple energy regimes, including the few-million-electron-volt fission neutron-energy region.

Unfortunately, measuring the amount of iridium-193m produced in a nuclear test was also uniquely difficult. It must be done by measuring the decay of the radioactive isomer, but the decay proceeds through two competing processes, gamma-ray emission and internal conversion, and the latter is very difficult to separate from the background[5]. The problem was first solved by some of the great figures from the radchem past of Los Alamos, such as Jim Gilmore, Don Barr, and Moses Attrep. The experimental problem was so difficult that other laboratories, such as Lawrence Livermore

[4] Radiochemistry measures only time-integrated quantities because its measurements reveal the cumulative result of a long, complex sequence of production-destruction reactions on the nuclei.

[5] Internal conversion is a nuclear decay process in which the nucleus changes to a lower energy level and maintains energy conservation by emitting an electron from an atomic shell. Because it is charged, that electron tends to be stopped in the sample, emitting x-rays as it slows down. Often those x-rays can be very difficult to separate from the background. In the more usual decay process, the nucleus emits a readily detected gamma ray as it decays to a lower energy level.

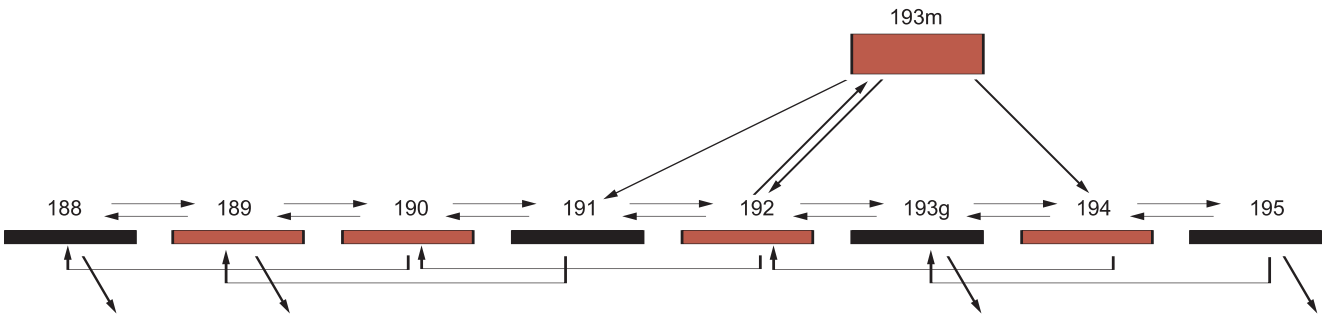


Figure 8. Reaction Pathways for Neutrons Hitting an Iridium Target
The different arrows correspond to neutron-induced reactions on iridium nuclei such as (*n*,), (*n*,*n*′), (*n*,2*n*) and (*n*,3*n*), where the left entry indicates the incident particle and the right entry indicates the outgoing particles. By measuring the various production rates of the radioactive isotopes iridium-189, -190, 192, -193m, and -194 when exposed to a particular neutron fluence, one can learn precious information on the cross sections for each reaction present in this reaction network. In particular, the inelastic neutron scattering reaction cross section for iridium-193 (*n*,*n*′) iridium-193m reaction cross section is most sensitive to neutrons in the few-million-electron-volt energy range and therefore can contribute to assessing the neutron fluence in this neutron energy range.

National Laboratory (Lawrence Livermore) and the Atomic Weapons Establishment in Great Britain, relied upon Los Alamos radiochemistry for this task.

Nuclear Cross Sections and Uncertainty Quantification. As mentioned earlier, to accurately infer neutron fluences from radiochemical measurements of isotopes after a nuclear test, it is not enough to determine the relative amounts of the various isotopes. The nuclear cross sections for producing those isotopes must also be known accurately. This has not been the case for iridium cross sections. In particular, the (n, n') neutron-scattering cross section that determines the production of the isomer iridium-193m is extremely difficult to measure because there are many different pathways leading to isomer production and some of them—for example, direct population of the isomer state through neutron scattering and internal conversion—cannot be observed. Figure 9 shows a diagram of the energy levels of the iridium nucleus and the many pathways that lead to population of the isomeric state.

Until recently, the only experimental data on iridium isomer production were obtained at incident neutron energies above 7.5 MeV by the Los Alamos radchem group mentioned (Bayhurst et al. 1975). Consequently, the historic isomer production cross-section data set used at Los Alamos for the last two decades was based almost exclusively on the nuclear-theory predictions of Ed Arthur of the Theoretical (T) Division at Los Alamos.

In the last few years, a collaboration between experimentalists at the Los Alamos Nuclear Science Center (LANSCE) and theoreticians in T-Division has determined and evaluated new data for the isomer-production cross section. LANSCE's GEANIE gamma-ray detector (see Figure 10) was used to measure the cascade of gamma rays that results when the excited iridium-193 nucleus loses energy on its way to populating the metastable isomeric state. The GEANIE measurements were undertaken by a Los Alamos–Lawrence Livermore collaboration involving Ron Nelson, Nick Fotiadis, Matt Devlin, John Becker, Paul Garrett, and Lee Bernstein. But GEANIE could not measure the contributions to the isomer production from processes that do not involve gamma rays. Those contributions had to be predicted from theory. Theory was also needed to predict certain gamma-ray feeding transitions that could not be measured directly because of experimental limitations.

The authors accomplished that task by incorporating advanced nuclear-reaction-theory models into the GNASH code, which was developed in T-Division for predicting nuclear cross sections. Those advanced models describe compound nucleus, pre-equilibrium, and direct mechanisms for a nucleus to reach an isomeric state. In order to accurately model isomer production, we had to understand the following nuclear properties: (1) optical potentials that describe the motion of the incoming and outgoing neutrons relative to the target nuclei, (2) the nuclear structure and decay properties of the low-lying levels (obtained from experiment) and highly excited levels (obtained from statistical theories of excited nuclei), and (3) the angular momentum transfer processes associated with the pre-equilibrium and compound nucleus decay mechanisms. Our

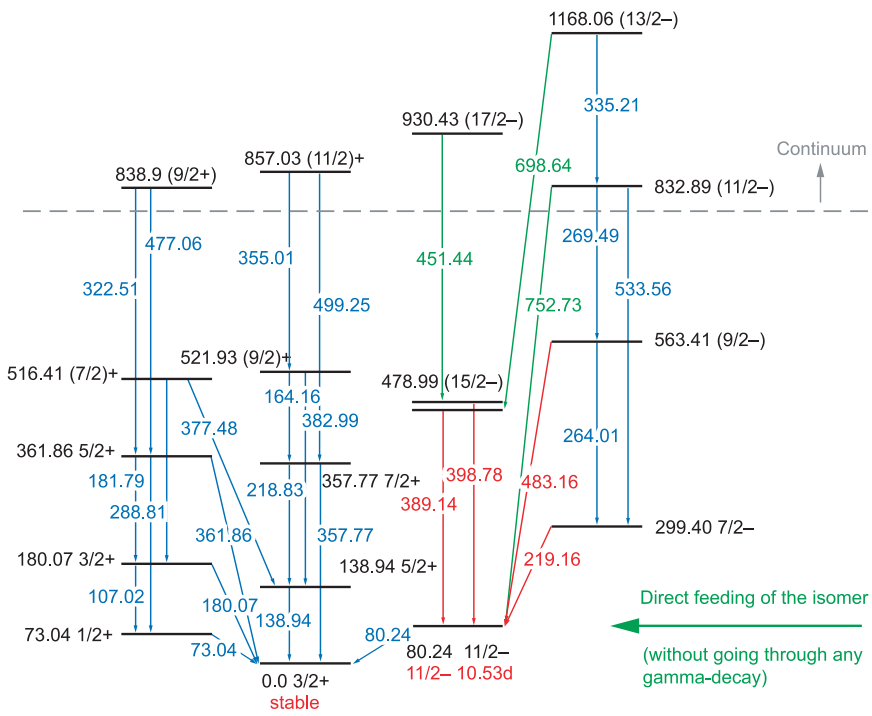


Figure 9. Pathways for Producing the Isomer Iridium-193m
This nuclear energy-level diagram shows the various pathways for producing the long-lived isomer state at an energy of 80 keV above the ground state. The GEANIE experiment clearly resolved the four strongest γ-ray transitions (red lines) that feed the 80-keV isomer. GNASH calculations were benchmarked against the GEANIE data for the strengths of those four transitions and then were used to calculate all other unaccounted for contributions to the isomer production cross section. The latter include the direct feeding of the isomer by neutron inelastic scattering (without going through the γ-ray cascade) and the other γ-ray transitions (green lines) that either reach the isomer or feed levels that reach the isomer.

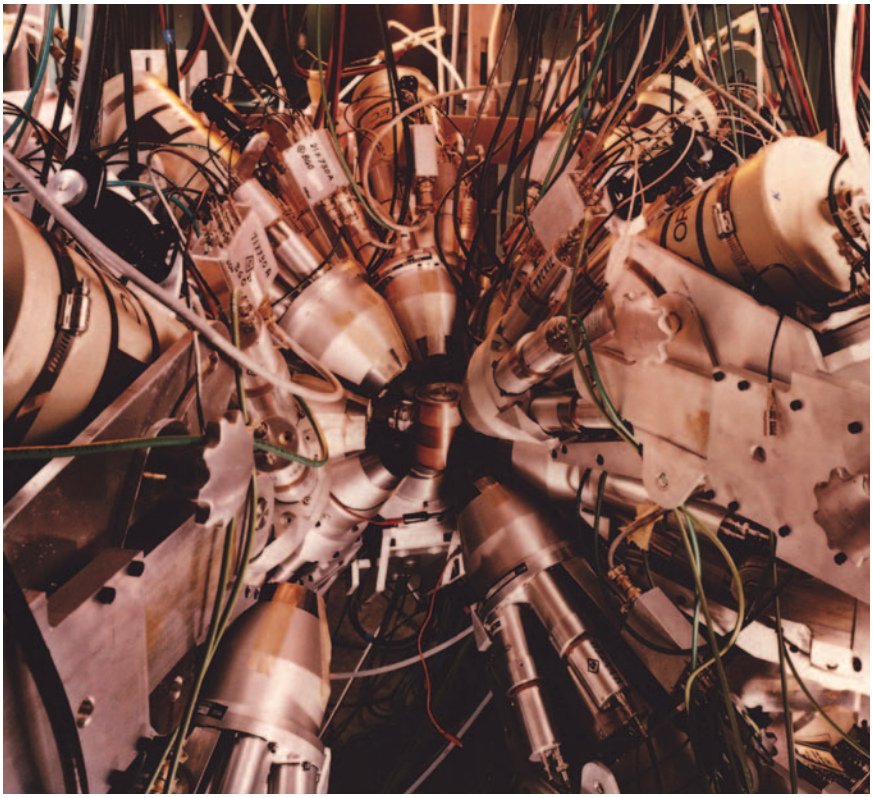


Figure 10. The GEANIE Detector
GEANIE (germanium array for neutron-induced excitations) is a 4π high-resolution γ-ray spectrometer installed at LANSCE's Weapons Neutron Research Facility. It can detect γ-rays from about 20 keV up to 8 MeV. The neutrons hitting the target samples cover the energy range from below 1 MeV to more than 200 MeV. The time-of-flight technique is used to determine precisely the energy of the incident neutrons, with a 22-m flight path. The GEANIE spectrometer was used to study details of the γ-ray cascade following the inelastic neutron scattering on iridium-193.

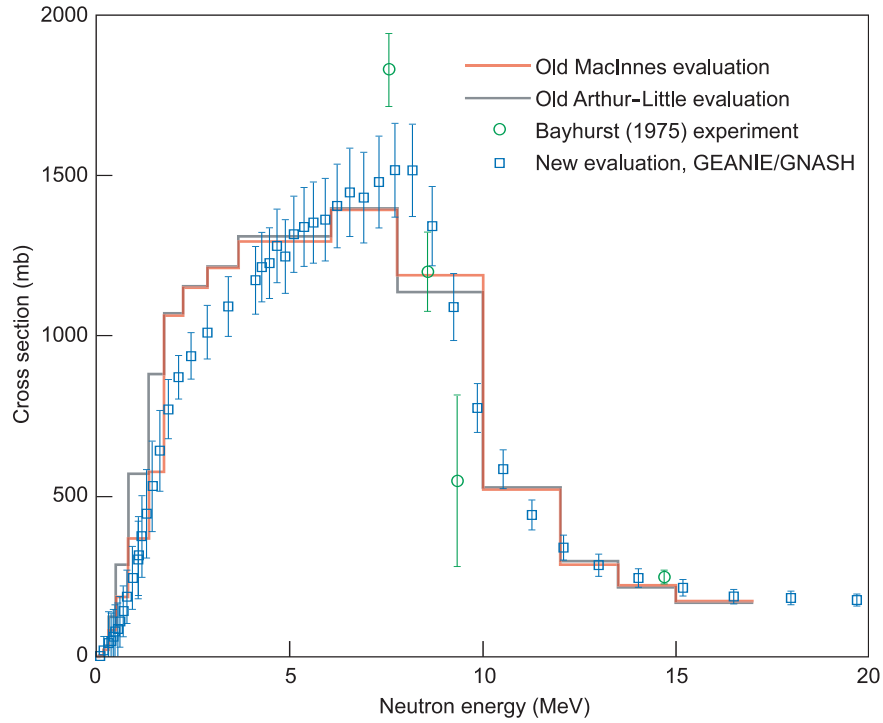


Figure 11. New Evaluated Production Cross Section for Iridium-193m
The new GEANIE/GNASH prediction for the 80-keV isomer production cross section in iridium-193 is shown here, covering the incident neutron energy range from the reaction threshold (80 keV) up to 20 MeV. The 1-σ standard deviations that come from uncertainties in both GEANIE data and GNASH reaction modeling are also plotted. This new cross section is compared with the historic one from T-Division (by Ed Arthur and Robert Little) that has been used until now in weapon physics work at Los Alamos. Note that our new result is in much better agreement with the MacInnes ad hoc fix to the Arthur-Little evaluation near threshold, which was incorporated to improve the agreement with data from critical assemblies.

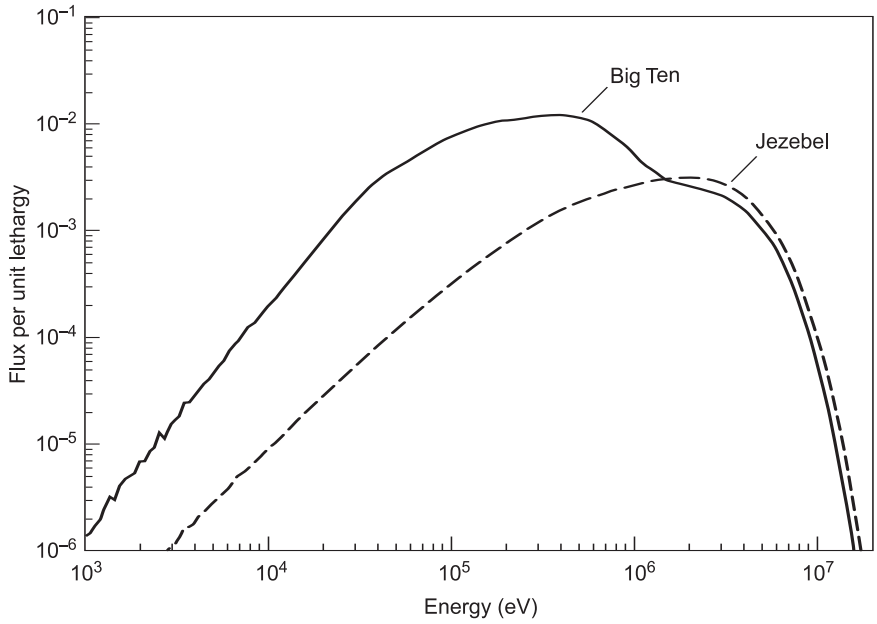


Figure 12. Hard and Soft Neutron Spectra
Assemblies with the highest average neutron energy are said to have the "hardest" spectra, whereas those that produce neutrons with a lower mean energy are said to have "softer" spectra. For example, the center of the Jezebel assembly (a sphere of plutonium) has one of the hardest spectra available. The Big Ten assembly, which has large amounts of uranium-238 and -235, has a much softer neutron spectrum.

research on these properties for iridium, together with extensive experience we have built up in analyzing similar data for other nuclei measured at LANSCE, allowed us to predict the various contributions to iridium isomer production using our advanced version of the GNASH code.

To test the accuracy of our calculational ability, we compared our GNASH cross-section predictions for the measured gamma-ray decay transitions with those determined from the GEANIE measurements. After we validated our predictive capability, we could apply the theory to predicting the unmeasured contributions with confidence. We could then evaluate the isomer-production cross section and its uncertainty using both the GEANIE and GNASH results.

Figure 11 shows our newly evaluated cross section for isomer production. The new GEANIE-GNASH

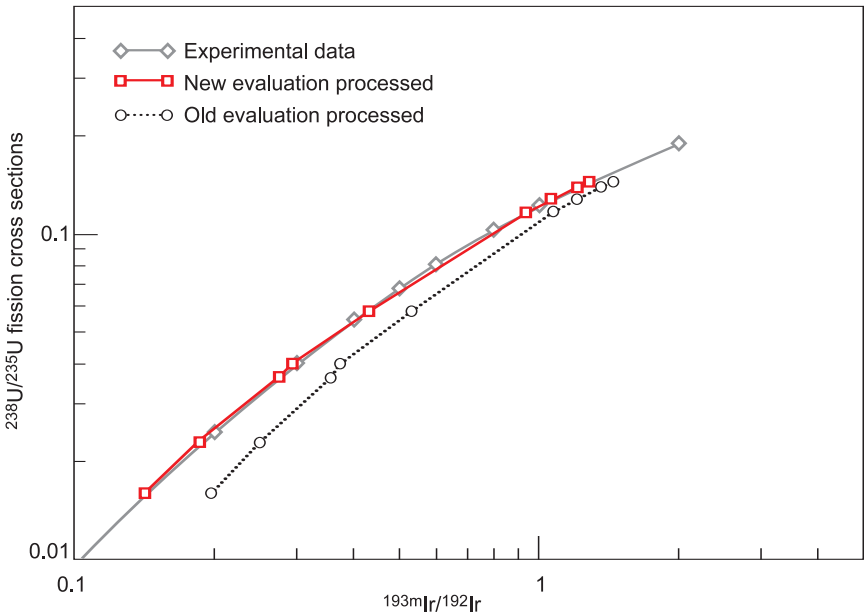


Figure 13. Iridium-193m Production Cross Section
Experimental data obtained with critical assemblies at LACEF were used to validate our new evaluation work. This figure represents the ratio of the iridium-193m production cross section to the production of iridium-192 (mainly through the neutron capture cross section of iridium-191) as a function of the ratio of uranium-238 to uranium-235 neutron-induced fission cross sections. The latter ratio represents the "hardness" of the neutron spectrum. This quantity changes with the location of the target in the critical assembly. Near the center, the neutron spectrum is quite hard; at larger distances, it softens. The slope of the experimental curve in this figure is therefore an indicator of the shape of isomer production cross-section, in particular near the threshold energy. Our new evaluation represents a net improvement over the older existing evaluation.

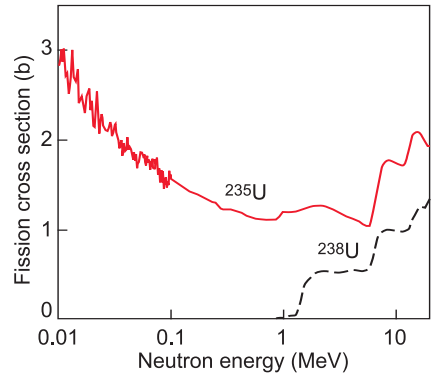


Figure 14. Fission Cross Sections for Uranium-235 and Uranium-238
Note the rapid increase of the uranium-238 fission cross section around an incident neutron energy of about 1 MeV. This threshold behavior causes the ratio of the uranium-238 to uranium-235 cross sections in critical-assembly experiments to increase with spectral hardness.

results cover the whole energy range of interest, from the threshold of the reaction at 80 kilo-electron-volts (keV) to above 20 MeV. Ed Arthur's old theoretical evaluation is also shown. Although the two results are similar overall, they also differ in subtle but important ways. In particular, our new cross section rises from threshold in a different manner, with a steeper slope. This outcome has important consequences, as will be described in more detail below. The uncertainties that we have derived for this cross section are shown as 1-σ error bars in Figure 11. The uncertainties that we have deduced include systematic and statistical errors, and they are associated with both the measured data and the GNASH nuclear model calculations.

Integral Data Testing at Critical Assemblies. With our new cross sections in hand, we will undertake weapon code simulations of specific past underground nuclear tests in which the nuclear devices were loaded with iridium radchem detectors and combine the calculated neutron fluences and our new cross sections to predict the iridium isotopic ratios produced in those tests. Because we have determined cross section uncertainties for the iridium reactions, we will also be able to provide uncertainties on the weapons code predictions of the iridium isotopic ratios. We will then compare the predicted ratios against actual post-test radchem measurements from those tests. Finally, we will work with designers to incorporate the results of those comparisons into the baseline certification calculations. Clearly, detailed aspects of this work will remain classified.

Interestingly, we have been able to validate our new iridium cross sections against an old but fascinating and unclassified set of iridium radchem data. Those data, obtained from fast critical-assembly experiments conducted over several decades at LACEF, provide a valuable integral test of our iridium cross sections. The fast critical assemblies at Los Alamos involve macroscopic quantities of special nuclear materials—plutonium and uranium-235 and -238—often in spherical configurations. When the critical mass is assembled, a self-sustaining chain reaction occurs, creating a neutron flux that has the energy spectrum typical of a fast fission-chain reaction. During the iridium radchem experiments, the flux of neutrons irradiated iridium foils placed inside the assembly, and the ratios of various iridium isotopes produced during irradiation were subsequently measured. One such ratio was irid-

ium-193m/iridium-192, in which the isomer came from the iridium-193 (n,n') reaction and the isotope, from the iridium-191 (n,γ) reaction. (Contributions from the iridium-193 ($n,2n$) reaction are very small in a critical assembly.)

Those old measured ratios can be compared with new predictions for these ratios obtained with our new cross sections. To predict the isotopic ratios, we must first predict the neutron energy spectrum of the fast critical assembly experiments using an MCNP radiation transport simulation and then fold that spectrum together with our iridium cross sections.

Clearly, fast critical assemblies provide valuable integral experiments to test our iridium cross sections because the neutron energy spectrum created in a fast critical assembly is skewed toward neutron energies that the iridium-193m diagnostic was developed to detect, that is, energies in the few-million-electron-volt region. But we also wanted to validate our cross sections over energies extending down to the threshold for isomer production, which is 80 keV. Again, iridium radchem data from old critical-assembly experiments have been invaluable. The various assemblies provide neutron energy spectra with varying average energies, depending on the critical assembly and the location within that assembly (see Figure 12). Fortunately for our iridium work, radiochemists had already conducted experiments in which iridium foils were loaded at various radial locations throughout a "traverse" of the Flattop assemblies (a core of HEU or plutonium surrounded by uranium-238). Those experiments involved neutron spectra ranging from "hard" (at the center of the assembly) to "soft" (at maximum distance from the center).

Figure 13 shows the radiochemical results obtained with the Flattop assembly for the isotopic ratio of iridium-193m to iridium-192 as a function of the hardness of the critical-assembly neutron spectra, where the spectral hardness is represented by the ratio of uranium-238 to uranium-235 fission cross sections. That fission cross-section ratio is used for two reasons: It increases with spectral hardness, or average neutron energy (see Figure 14), and it can be measured within the assembly, at the very spot where the iridium foils have been placed. Figure 13 also shows our calculated results for the iridium isotopic ratios, as well as results from Ed Arthur's old evaluation. The good agreement between measured data and our new calculated results validates our iridium-193m (n,n') cross section in the few-million-electron-volt region. Moreover, our reproduction of the shape of the experimental curve derived from the Flattop integral experiments validates the shape of the new GEANIE/GNASH microscopic cross section in Figure 11 as it rises from threshold.

The validation of the new isomer production cross section near threshold represents a breakthrough. Several years ago, Mike MacInnes of Los Alamos first undertook calculations of the Flattop critical assembly data in Figure 13 with the historic Ed Arthur's iridium-193 (n,n') isomer cross section used at Los Alamos at the time. He noted that the calculated shape did not agree well with the measured shape. This observation led him to make a change to the shape of the historic cross section near threshold. Our new result for this same cross section, based on independent LANSCE data and nuclear model calculations, has confirmed MacInnes' intuition.

Conclusions

Our ability to predict important nuclear cross sections and quantify uncertainties in those predictions has advanced considerably in the last decade. The rates of neutron-induced fission reactions are crucial to the performance of weapons. That is why reducing the uncertainty in those rates leads to more confident predictions using the Los Alamos weapons simulation codes. In addition, increased accuracy of neutron-scattering results obtained with radchem tracers has contributed to better assessments of past nuclear tests. ■

Acknowledgment

The authors would like to thank Dr. Kenneth Hanson from the Continuum Dynamics Group at Los Alamos for his insight into Peelle's Pertinent Puzzle (discussed in this article) and, more broadly, for stimulating discussions on modern statistical analysis techniques.

Further Reading

Bayes, T. 1763. An Essay Towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philos. Trans. R. Soc. London* **53**: 370.

Bayhurst, B. P., G. S. Gilmore, R. J. Prestwood, J. B. Wilhelmy, N. Jarmie, B. H. Erkkila, and R. A. Hardekopf. 1975. Cross Sections for *(n,xn)* Reactions between 7.5 and 28 MeV. *Phys. Rev.* C **12**: 451.

Becquerel, H. 1896. Emission de Radiations Nouvelles par l'Uranium Metallique. *C. R. Acad. Sci.* **122**: 1086.

Bohr, N., and J. A. Wheeler. 1939. The Mechanism of Nuclear Fission. *Phys. Rev.* **56**: 426.

Chadwick, J. 1932. Possible Existence of a Neutron. *Nature* **129**: 312.

Hahn, O., and F. Strassman. 1939. Concerning the Existence of Alkaline Earth Metals Resulting from Neutron Irradiation of Uranium. *Naturwissenschaften* **27**: 11.

Hayes, B. 2000. Dividing the Continent. *Am. Scient.* **88** (6): 481.

Mamdouh, A., J. M. Pearson, M. Rayet, and F. Tondeur. 1998. Large-Scale Fission-Barrier Calculations with the ETFSI Method. *Nucl. Phys. A* **644** (4): 389.

Meitner, L., and 0. Frisch. 1939. Disintegration of Uranium by Neutrons: A New Type of Nuclear Reaction. *Nature* **143**: 239.

Möller, P., and A. Iwamoto. 2000. Realistic Fission Saddle-Point Shapes. *Phys. Rev. C* **61**: 047602.

Möller, P., and J. R. Nix. 1981a. Atomic Masses and Nuclear Ground-State Deformations Calculated with a New Macroscopic-Microscopic Model. *AT. Data Nucl. Data Tables* **26** (2): 165.

———. 1981b. Nuclear Mass Formula with a Yukawa-Plus-Exponential Macroscopic Model and a Folded-Yukawa Single-Particle Potential. *Nucl. Phys. A* **361** (1): 117.

Möller, P., and S. G. Nilsson. 1970. The Fission Barrier and Odd-Multipole Shape Distortions. *Phys. Lett. B* **31** (5): 283.

Möller, P., A. J. Sierk, and A. Iwamoto. 2004. Five-Dimensional Fission-Barrier Calculations from ^{70}Se to ^{252}Cf. *Phys. Rev. Lett.* **92** (7): 072501.

Möller, P., J. R. Nix, and K.-L. Kratz. 1997. Nuclear Properties for Astrophysical and Radioactive-Ion-Beam Applications. *AT. Data Nucl. Data Tables* **66** (2): 131.

Möller, P., D. G. Madland, A. J. Sierk, and A. Iwamoto. 2001. Nuclear Fission Modes and Fragment Mass Asymmetries in a Five-Dimensional Deformation Space. *Nature* **409** (6822): 785.

Möller, P., J. R. Nix, W. D. Myers, and W. J. Swiatecki. 1995. Nuclear Ground-State Masses and Deformations. *AT. Data Nucl. Data Tables* **59** (2): 185.

Nix, J. R. 1972. Calculation of Fission Barriers for Heavy and Superheavy Nuclei. *Annu. Rev. Nucl. Sci.* **22**: 65.

Pauli, H. C. 1973. On the Shell Model and its Application to the Deformation Energy of Heavy Nuclei. *Phys. Rep.* **7** (2): 35.

Petrzhak, K. A., and G. N. Flerov. 1940. Uber die Spontane Teilung von Uran. *C. R. Acad. Sci. USSR* **28** (6): 500.

Rutherford, E. 1911. The Scattering of α and β Particles by Matter and the Structure of the Atom. *Philos. Mag.* **21** (6): 669.

Seaborg, G. T., E. M. McMillan, J. W. Kennedy, and A. C. Wahl. 1946. Radioactive Element 94 from Deuterons on Uranium. *Phys. Rev.* **69** (7–8): 366.

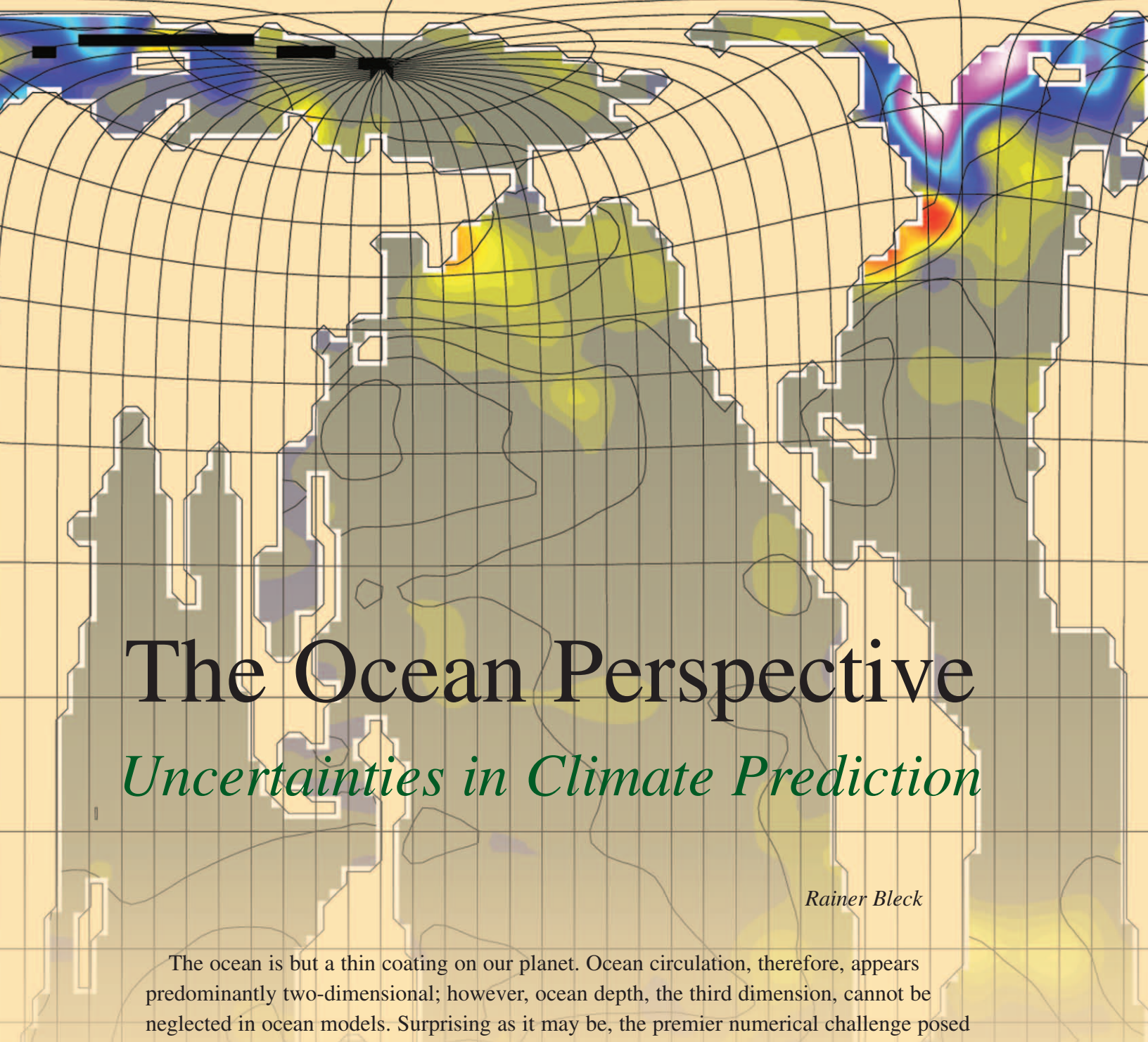
Shcherbakov, O., A. Donets, A. Evdokimov, A. Fomichev, T. Fukahori, A. Hasegawa, et al. 2002. Neutron-Induced Fission of ^{233}U, ^{238}U, ^{232}Th, ^{239}Pu, ^{237}Np, ^{nat}Pb, and ^{209}Bi Relative to ^{235}U in the Energy Range 1–200 MeV. *J. Nucl. Sci. Technol.* **1** (2): 230.

Staples, P., and K. Morley. 1998. Neutron-Induced Fission Cross-Section Ratios for ^{239}Pu, ^{240}Pu, ^{242}Pu, and ^{244}Pu Relative to ^{235}U from 0.5 to 400 MeV. *Nucl. Sci. Eng.* **129** (2): 149.

Strutinsky, V. M. 1968. "Shells" in Deformed Nuclei. *Nucl. Phys. A* **122**: 1.

———. 1967. Shell Effects in Nuclear Masses and Deformation Energies. *Nucl. Phys. A* **95**: 420.

For further information, contact Mark B. Chadwick (505) 667-9877 (mbchadwick@lanl.gov).



The Ocean Perspective

Uncertainties in Climate Prediction

Rainer Bleck

The ocean is but a thin coating on our planet. Ocean circulation, therefore, appears predominantly two-dimensional; however, ocean depth, the third dimension, cannot be neglected in ocean models. Surprising as it may be, the premier numerical challenge posed for ocean models used for climate prediction is keeping the warm poleward-flowing surface water thermally insulated from the cold abyssal return flow—as insulated as it is in nature. Los Alamos supports several approaches to ocean simulations, whose results give a hint about the uncertainties involved in climate prediction. The model designed to come closest to preserving the warm poleward and cold return flows of the ocean "conveyor" is the layer model, which pictures the ocean as a stack of immiscible layers. Compared with other models, the layer model also produces more stable oceanic circulation in the face of climate changes. Yet the jury is still out on whether "more stable" is the same as "more realistic."

Is it preposterous to predict Earth's climate 50 or 100 years ahead if we cannot reliably forecast the weather two or three days into the future? Fortunately, the situation is not as hopeless as one may think. There are fundamental differences between the two tasks.

Mathematicians classify weather prediction as an "initial value" problem because the accuracy of a weather forecast depends crucially on how well the initial state of the atmosphere is known. Climate prediction, on the other hand, is primarily a "boundary value" problem. In this case, the main task is to reproduce the time-averaged flow of solar energy through the nooks and crannies of the land-ocean-atmosphere system. To do so well, one needs to know those nooks and crannies, and one needs to know how much energy arrives at the top of the atmosphere as a function of latitude and time of year. But the exact locations of the transient disturbances that determine the oceanic and atmospheric "weather" need not be known, either initially or at a later time. In essence, when we predict future climates, we try to assess whether modifying certain parameters, such as the ellipticity of the earth's orbit or the chemical composition of the atmosphere, will change the way energy flows through the earth system. This task does not critically depend upon our ability to predict tomorrow's weather or the onset of the next El Niño—even though a forecast model that does well in these respects will increase our confidence in the correctness of the climate forecast.

Simulating systems that are as complex as Earth's climate is hard. Two types of errors may affect the simulation: errors in the physics of the model and errors in the mathematical approximations needed to simulate climate processes on a computer. Being able to distinguish between these two error types may help us develop more-accurate climate models. But to separate errors, scientists need tools, and model diversity is among the few available ones. In the realm of ocean modeling, Los Alamos has been supporting model diversity for over a decade. Several ocean-circulation models have been brought to or developed at the Laboratory, and they are designed to solve the same physical problem while being numerically dissimilar. By comparing their results, scientists get a feel for the size of the uncertainties. This article will use three examples related to Elf Niño, the heat-carrying ocean conveyor, and oceanic carbon sequestration to illustrate this approach.

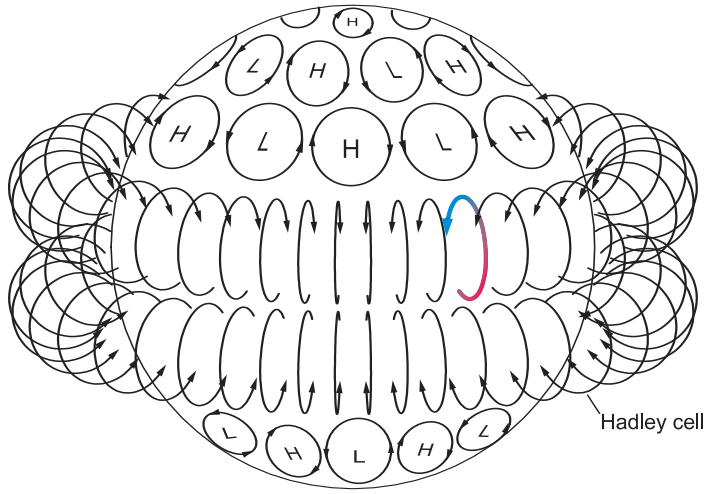


Figure 1. Heat Transport in the Atmosphere
This schematic view shows two atmospheric circulation modes important for poleward heat transport. A vertical-meridional overturning circulation (Hadley cell) dominates near the equator. Horizontally rotating eddies (the highs and lows on weather maps) dominate at mid to high latitudes.

Modes of Poleward Heat Transport

Our planet absorbs solar energy at low latitudes and radiates energy back into space at high latitudes. This is so because the earth is a sphere and its axis of rotation is more or less perpendicular to its orbital plane around the sun. For this system to remain in a steady state, heat on earth must continually flow poleward in both hemispheres. Transporting this heat is the job of the atmosphere and ocean because, in contrast to the solid earth, they can move heat efficiently by setting up warm currents flowing poleward and cold ones flowing back to the equator.

From here on things get complicated. The earth's rotation greatly inhibits meridional displacement of water or air because a northward- or southward-moving fluid parcel away from the equator also changes its distance from the earth's axis. In fact, the angular-momentum balance constraints resulting from the earth's rotation are so severe that the atmosphere can maintain a meridional overturning circulation (a closed loop consisting of air rising at low latitudes and sinking at high latitudes) only near the equator in the so-called Hadley cell (Figure 1). At mid-to-high latitudes, the earth's rotation forces the atmosphere to resort to a different mode of heat transport, namely, transient eddies, popularly known as highs and lows, which intermittently push warm air poleward and cold air equatorward over distances too small for the angu-

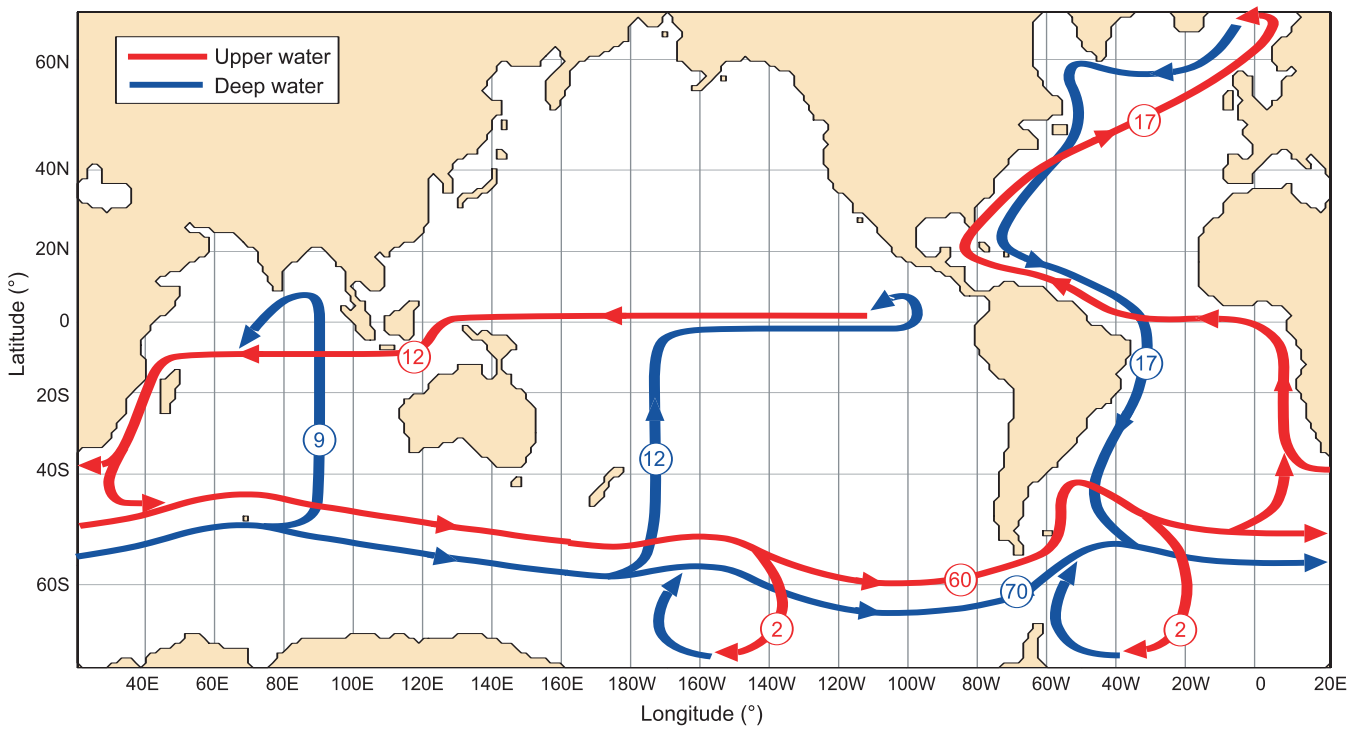


Figure 2. Heat Transport in the Ocean
The thermally forced ocean circulation spans ocean basins, as shown in this figure. Vertical and horizontal details are simplified but less so than in Broecker (1991). Wind-driven currents are omitted except for the Antarctic Circumpolar Current. Circled numbers represent transport in sverdrups (1 Sv = 106 m^3 s^{-1}, corresponding to roughly the volume transport of five Amazon Rivers). The schematic does not reflect the fact that downwelling takes place in geographically confined regions (Greenland/Norwegian Sea, Weddell Sea, and Ross Sea) while upwelling is a much more widespread process. Thus, not all the water entering the Indo-Pacific basins from the south up-wells in the specific locations indicated in the drawing. (Adapted from Sun and Bleck 2001a and Schmitz 1996).

lar momentum constraint to kick in.[1] The two modes of heat transport in the atmosphere and their respective geographic domains are depicted schematically in Figure 1.

In the ocean, in contrast to the atmosphere, steady meridional motion can be sustained over long distances when a current can "rub" against a continental margin and thereby shed momentum. This is why meridional ocean currents, such as the Gulf Stream, must always flow along the edge of an ocean basin, never in the middle. (Emphasis here is on the word "meridional." East-west currents can cross ocean basins in an unrestricted manner. Otherwise, the warm waters of the Gulf Stream would not be able to reach Europe.) Eddies, analogous to those in the atmosphere, do exist in the ocean, but their contribution to heat transport tends to be overshadowed by the contribution of the boundary currents. The Southern Ocean, being devoid of meridional land barriers, is the obvious exception; there, as in the atmosphere, ocean eddies play a primary role in heat transport.

The ability of the ocean to maintain steady meridional motion over considerable distances actually allows the ocean to develop two types of heat transport mechanisms not found in the atmosphere: a Hadley cell–like meridional overturning circulation extending all the way to the subpolar seas—dubbed the ocean conveyor (Broecker 1991)—and a basin-spanning horizontal gyrating motion. The former, depicted schematically in Figure 2, is primarily maintained by differential heating and cooling; the latter, by the torque exerted on the ocean by the prevailing pattern of tropical easterlies and extratropical westerlies.

[1] Even though extratropical eddies are as flat as pancakes, their flow field is not entirely two dimensional; in fact, they draw their energy from the rise/descent of warm/cold air masses. Their residual effect, if analyzed in a proper entropy-oriented framework, therefore, is to extend the Hadley cell to higher latitudes.

Implications for Ocean and Climate Modeling

To faithfully replicate the relevant heat-transport mechanisms on our planet, a climate model must be able to reproduce the action of atmospheric lows and highs without which there would be hardly any heat transport in the atmospheric submodel. In other words, the atmospheric submodel must be what ocean modelers refer to as "eddy resolving." In the oceanic submodel, on the other hand, the first order of business is to correctly simulate the major current systems, both those associated with the wind-driven horizontal gyre circulation and those associated with the thermally driven meridional overturning circulation.

This is not to say that the effect of ocean eddies can safely be neglected. Wherever they are in the ocean (and they are almost everywhere), eddies will transport some heat. However, in most oceans, except the Southern Ocean, the contribution of the eddies is overshadowed by the contribution of meridional current systems. As a result, the penalty for "parameterizing" the eddies' role, instead of explicitly resolving the eddies, is minor. Turning this argument around, one should expect the Southern Ocean to emerge as a major Achilles' heel in noneddy-resolving ocean modeling.

Eddy resolution in the ocean is a major problem. According to hydrodynamic instability theory, tailored to fluid motion on a rotating sphere, eddy size depends on the vertical density contrast in the fluid. Because this contrast is much smaller in the ocean than in the atmosphere, ocean eddies turn out to be roughly 10 times smaller in diameter (that is, 100 times smaller in area) than their atmospheric counterparts. Hence, the number of eddies to be tracked by an eddy-resolving ocean model through their individual life cycles exceeds by two orders of magnitude the number of eddies in a global weather model. Furthermore, in the context of climate, individual eddies would have to be simulated not only for the duration of a 5- or 10-day weather forecast, but also for decades or possibly centuries. This task is beyond the capabilities of even our biggest and fastest computers.

Contrary to common perception, the ocean is quite shallow, a thin coating on our planet, and oceanic circulation appears, therefore, predominantly two-dimensional. However, the presence of meridional overturning circulations and the concomitant reversal of current direction with depth mean that the third (vertical) dimension cannot totally be neglected when modeling the ocean. Surprising as it may sound, the premier numerical challenge posed by the third dimension in ocean models used for climate prediction is to keep the warm poleward-flowing surface water thermally insulated from the cold abyssal return flow—as insulated as it is in nature. Given the long time scales involved (decades to centuries) and the relative proximity of the two circulation branches (a few kilometers), this is indeed a major challenge. It has motivated the development of a class of ocean models that, instead of carrying ocean state variables on a rigid, crystal-like lattice, picture the ocean as a stack of immiscible layers whose thicknesses are allowed to evolve freely in space and time. By allowing grid cell interfaces (and the state variables riding on them) to bob up and down with the vertical component of motion, these so-called layer models control vertical mixing processes much better than models based on a rigid spatial grid. (The dispersive effect of an oscillating vertical motion field on such properties as temperature in a fixed-grid ocean model is illustrated in Figure 3). As a result, warm surface currents in a layer model are less likely to lose heat through contact with the cold return flow than those in a traditional fixed-grid ("level") model. In theory, at least, this difference translates into a more robust heat-delivery system and a more accurately simulated climate.

Potential density, defined as density corrected for compressibility effects, is a proxy for entropy in seawater and hence is conserved in the absence of heat-transferring, or diabatic, processes. Because oceanic flow below the surface layer generally comes close to being adiabatic, the layers in a layer ocean model are typically chosen to coincide with constant potential-density, or isopycnic, layers. The resulting impermeability of layer interfaces under adiabatic flow conditions allows vertical property exchange by diabatic mixing, to the extent that it occurs, to be modeled explicitly before a background of zero numerical mixing.

Replacing the traditional Eulerian vertical coordinate by a Lagrangian one, tied to the oceanic potential density field, sounds easier than it is. Given the small but persistent background mixing in the ocean, maintenance of a steady climate state requires that each parcel of seawater communicate with the atmosphere at least intermittently to replenish its temperature and salinity—the two ingredients that set the density of seawater. This is to say that each layer in a layer model must be allowed to "outcrop," or rise to the surface. Picturing the world ocean as a lens of light, warm water centered on the equator and floating on a body of dense, cold water, one readily sees that the densest layers outcrop closest to the poles, layers of intermediate density outcrop at mid-latitudes, and so forth (refer to Figure 4). To avoid having to deal with time-dependent lateral boundaries for individual coordinate layers, today's isopycnic models extend ocean layers, regardless of the actual extent of the water, over the

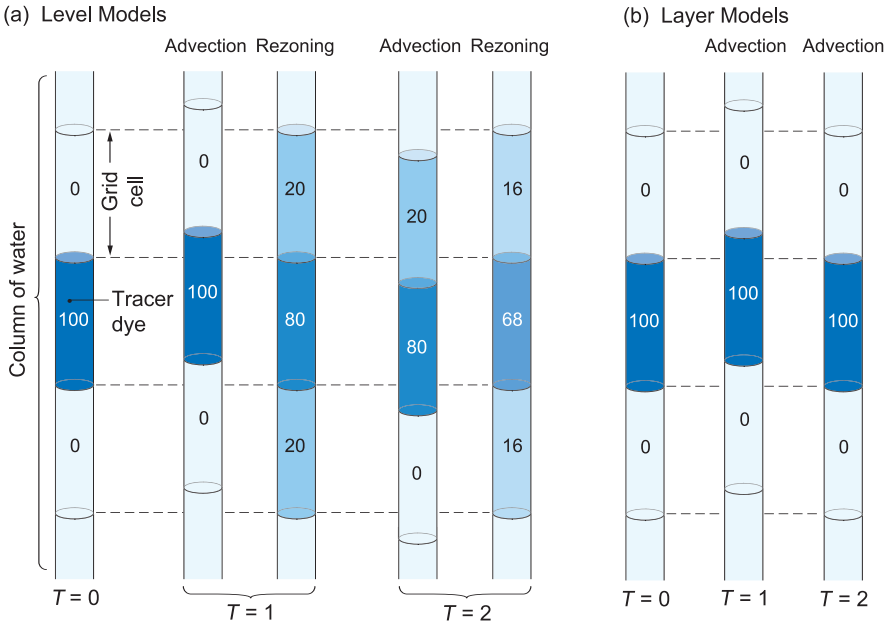


Figure 3. Anomalous Vertical Mixing in Fixed-Grid Models
(a) This schematic illustrates numerical dispersion in a water column, resulting from oscillatory vertical motion typically associated with passing gravity waves. Time increases from left to right. Shown is a vertical stack of three grid cells. The initial state, *T* = 0, is chosen to coincide with the wave trough, at which time the center grid cell is assumed to be filled with a tracer of concentration 100. In the advection step at *T* = 1, the approaching wave crest causes the water in all three cells to rise by a distance corresponding to one-fifth of the vertical cell size. The clock is stopped momentarily to allow the tracer to be reapportioned, or rezoned, among the original grid cells, which in contrast to the water column, stay fixed in a level ocean model. Because of rezoning, the tracer is split between two cells (*T* = 1). Next, the clock is running again. The approaching next wave trough causes the water column to return to its initial position during advection, at *T* = 2. With the clock stopped again, the tracer is being rezoned a second time. Tracer concentration in the center cell is now down to 68, with the remainder spread over the two adjacent cells. Note that this is an extreme example. Dispersion can be reduced by use of more sophisticated rezoning schemes. Also, gravity waves, while ubiquitous, usually have smaller amplitude than assumed here. (b) It is important to note that layer models skip the rezoning steps and thereby maintain a concentration of 100 in the center cell.

whole model domain as empty or massless layers. All these conditions translate into tricky numerical issues, making layer models inherently more complex than traditional level models.

Because of these tradeoffs, neither model class can be regarded as superior in every respect in simulating the global ocean circulation. However, two models that start from the same physics—including the "closure" model that approximates the effect of turbulent exchange processes at the small, unresolved scales—but express that physics in different mathematical form provide important insight into the inevitable degradation inherent in solving differential equations by computer. This comparative approach, therefore, affords some measure of the overall uncertainties in climate prediction.

It is important to note that the two ocean-model classes differ not only in their numerical representation of a given set of differential equations but also in the differential equations themselves. This begs the question, "how can there be two sets of equations for a single, uniquely defined physical problem?" The answer is that the underlying physical principles (Newton's law, conservation of mass, and others) can be cast in different forms, depending on which variables in the set consisting of depth, temperature, salinity, density, and velocity are treated as dependent variables. In level models, depth is an independent variable, whereas water density is a dependent variable, stepped forward in time as one solves prognostic equations for temperature and salinity. The equations governing layer models, on the other hand, treat density as an independent variable and, in the spirit of maintaining consistency between the number of unknowns and equations, they treat depth (in the form of layer thickness) as a dependent variable. It is this switch, rather than variations in the way differential equations are translated into algebraic ones, that gives different properties to the solutions obtained from level and layer models.

Los Alamos Contributions

In the early 1990s, the Department of Energy (DOE) Office of Science joined other federal agencies in funding the development of layer ocean models for climate prediction. The main reason was the perceived need to enrich the ocean model "gene pool," which at that time was rather sparse and showed signs of model inbreeding. Today, both level and layer models are firmly established at Los Alamos. The level model class is represented by the Los Alamos–developed Parallel Ocean Program (POP). For a detailed account of ocean-modeling advances achieved through

development of POP, refer to Malone et al. (1993, 2003). The layer model class is represented by the Miami Isopycnic Coordinate Ocean Model (MICOM) by Bleck et al. (1992) and its hybrid-coordinate offshoot HYCOM by Bleck (2002).

Hybrid-coordinate models are designed to combine the advantages of layer and level models. Starting at the surface, one assigns progressively larger "target" potential-density values to coordinate layers in hybrid models. Each coordinate layer is expected to track its assigned isopycnic layer in the model domain in space and time but may deviate from it to form a conventional constant-depth layer if (and only if) the target density is too low to exist in a given water column. Layers assigned to relatively warm, or low-density, water, which in traditional isopycnic models would only exist at low latitudes, thereby are allowed to molt into constant-depth layers poleward of their outcrop latitude. These redefined layers provide a framework for solving the model equations in subpolar oceans, where the lack of vertical density contrast makes it hard to represent vertical structure in terms of density classes.

Judging from the willingness of such federal agencies as the Naval Research Laboratory and the National Weather Service to adopt HYCOM (see, for example, http://www7320.nrlssc.navy.mil/ATLhycom1-12/skill.html), the hybrid model concept is widely being regarded as a significant step toward creating a flexible, multipurpose, next-generation ocean model. The COSIM (for Climate, Ocean, and Sea Ice Modeling) group at Los Alamos is under contract with the DOE Office of Science to produce a hybrid-coordinate version of POP as well.

The algorithm in HYCOM that determines whether a given coordinate layer can retain its isopycnic character at a given location or whether it must be assigned a constant thickness and be "frozen" in space has elements in common with the Los Alamos–developed arbitrary Lagrangian-Eulerian (ALE) technique (Hirt et al. 1974). However, whereas traditional ALE applications focus on maintaining a nonzero mesh size, the HYCOM algorithm addresses the more vexing problem of moving coordinate layers through the fluid to realign them with their respective target isopycnals after they have become separated. An illustration of how hybrid-coordinate models work in practice is given in Figure 5.

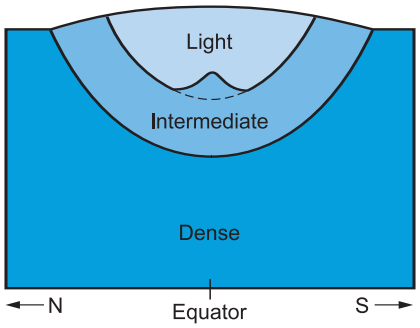


Figure 4. The Ocean as a Lens of Light Water Floating on Dense Water
As their densities increase, ocean layers outcrop progressively closer to the poles. Coordinate layers in MICOM follow the same general pattern.

Examples of Multimodel Climate Sensitivity Experiments

The vagaries of weather forecasts are the butt of jokes. Yet the meteorological community has rather precise information about the "skill" of numerical models used in daily forecasting and about the associated uncertainties. This information is precise because weather models are intended to duplicate the behavior of a readily observable system and because gathering statistical information about model skill is made easy by the large and ever-growing ensemble size.

The situation is quite different in decadal to centennial climate prediction because of the lack of verification data, the sheer number of natural processes contributing to the steadiness of climate (or its change, as the case may be), and the need to either treat in cursory fashion (parameterize) or totally omit from the model those processes that are deemed less central to the climate problem than others. Uncertainty quantification in long-range climate prediction, therefore, is a science that arguably is not even in its infancy.

Not much needs to be said about the lack of verification data. Important climate-relevant aspects of the earth system, such as atmospheric greenhouse-gas concentrations and the oceanic abyssal circulation, have been observed only in the last half century in sufficient detail to validate three-dimensional climate models. This observational record is vitally important as it provides a glimpse at the performance strengths and limitations of today's climate models, but it cannot serve as a database for rigorously assessing model skill. Stated differently, the 50-year observational record allows us to check the appropriateness of certain parameterizations (also referred to as physical closure assumptions) in climate models, but as an "ensemble" of one, it is insufficient for quantitatively evaluating prediction uncertainty.

At present, the focus in the climate research community is on the number (and ranking) of climate-contributing natural processes and on the need to parameterize. One can argue that, given the complexity of the climate problem and the finite nature of computing resources, there is not a single process that is not, in one way or another, parameterized in a climate model. The omission of possibly relevant detail begins with the transfor-

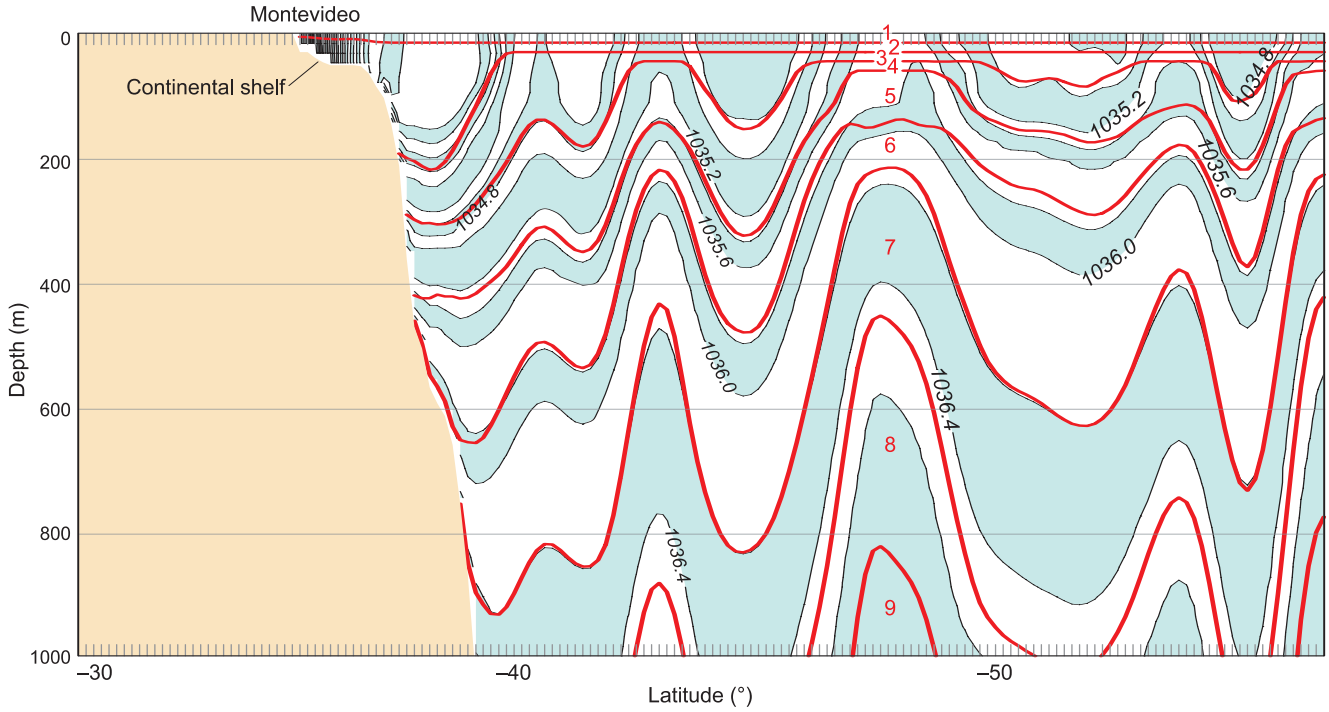


Figure 5. Ocean Density and Hybrid Coordinates Near the Falkland Islands

This is a sample vertical section through a HYCOM solution extending south from Montevideo into the eddy-rich confluence region of the Brazil and Falkland currents. The red numbers from 1 to 9 are the hybrid layers. The South American continent is shown at left. The latitude (°) is marked as negative numbers along the bottom. The heavy red lines represent HYCOM's coordinate surfaces; the shaded contours, outlined by light black lines, represent potential density in kilograms per cubic meter. Tick marks along the abscissa indicate grid resolution (approximately 15 km). The ordinate shows depth in meters. Note that coordinate surfaces follow isopycnals at depth but turn horizontal near the surface whereas the associated isopycnals outcrop. Density undulations indicate the presence of "cold-core" and "warm-core" eddies (which in the southern hemisphere spin clockwise and counterclockwise, respectively). Crowded isopycnals on the continental shelf indicate the presence of low-salinity Rio de la Plata water.

mation of the differential equations that govern the behavior of the natural system into computer-solvable algebraic equations. The truncation of the spectrum of scales at a chosen mesh size immediately divides processes into spatially resolved and unresolved ones, the latter requiring a physical closure assumption. A good example of a closure scheme for processes taking place on spatial scales too small to be resolved by a climate model is the wind-induced turbulent mixing below the sea surface. Since this turbulence stirs up water from depths of tens or even hundreds of meters, it strongly affects sea surface temperature. Disregarding or poorly parameterizing it, therefore, has dire consequences on the representation of air-sea exchange processes in our models.

Errors associated with the inevitably imperfect physical closure of unresolved processes are compounded by errors introduced by solving algebraic instead of differential equations; these so-called truncation or discretization errors mainly affect the resolved scales. Hence, climate forecasts are fraught with a mixture of physical closure errors and numerical truncation errors.

Notwithstanding efforts by groups such as the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory (http://www.pcmdi.llnl.gov), the climate community is still largely unable to separate the effects of physical and numerical errors on a climate forecast. One of the few tools at our disposal, as already mentioned, is developing multiple climate models that employ identical physical-closure schemes but are based on different numerics. This approach leads to the need for what was earlier referred to as genetic diversity in climate models. The differences between level and layer models arguably provide such diversity and hence open the door to experimentation aimed at separating physical from numerical model errors. A few examples of such experimentation are given below.

El Niño-like Variability in Climate Models. Much of the discussion about global warming focuses on

the question of whether the currently observed global temperature rise can be attributed to the inherent natural variability of the ocean-atmosphere system or whether it is a consequence of increased greenhouse gas concentrations. In order to clarify this question through numerical simulation, one obviously needs a climate model with a proven ability to simulate the multitude of ocean-atmosphere feedback mechanisms giving rise to natural variability.

The biggest observed climate variability on interannual time scales is associated with the so-called El Niño–Southern Oscillation (ENSO) phenomenon. The ocean-atmosphere system in the tropics is known to switch back and forth between two states, one of which (La Niña) is characterized by strong trade winds and strong upwelling of cold subsurface water in the equatorial eastern Pacific, whereas the other (El Niño) is characterized by weak trade winds and weak upwelling. Both states appear to be self-sustaining in the sense that strong/weak upwelling caused by strong/weak trade winds tends to support the underlying wind anomaly. A particular signal telling the coupled system to initiate the switch from one state to the other has not yet been identified. Efforts to predict that switch, therefore, have not advanced beyond the stage of what may euphemistically be described as early detection.

The ENSO coupled mode is often used as a yardstick for how well a climate model handles internal variability. Most coupled models are actually capable of producing an ENSO-like variability mode (AchutaRao and Sperber 2002), but a fair amount of parameter tuning is usually required before those models come close to simulating the observed amplitude, frequency, and spatial anomaly pattern of the genuine ENSO. Tuning attempts usually focus on the turbulence closure scheme for the oceanic and atmospheric boundary layers, but changing the scale selectivity of the model by modifying the computational mesh can also have a surprisingly strong effect.

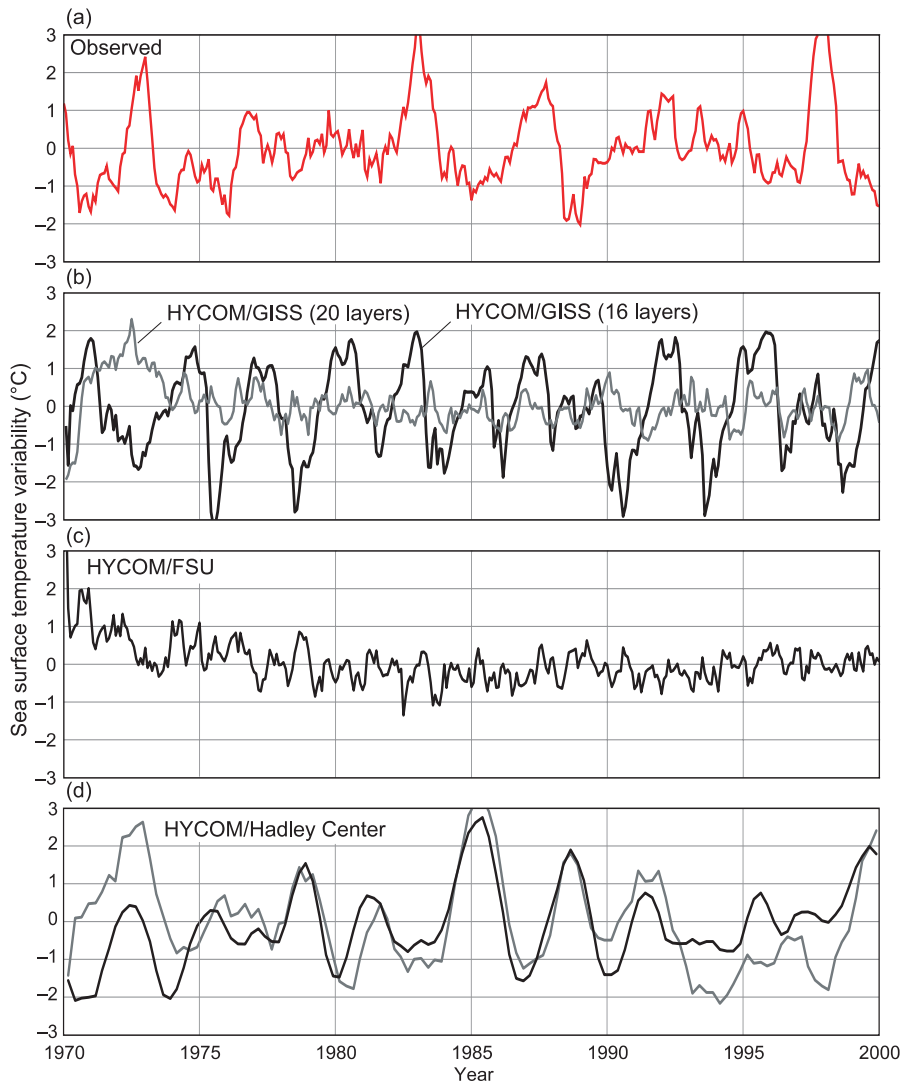


Figure 6. El Niño Variability in Climate Models

An observed 30-year time series of El Niño–related sea-surface temperature variability shown in (a) (Niño3 index, °C) is compared with corresponding time series obtained from three atmospheric circulation models, all of which have the oceanic component HYCOM in common: (b) model from the Goddard Institute for Space Studies (GISS) at NASA; (c) model from Florida State University (FSU); and (d) model from the Hadley Centre in the United Kingdom. Two curves within a panel indicate two runs based on different parameter choices: number of layers in (b) and different turbulence surface mixing in (d). The large model-to-model variation in Niño3 amplitude is largely unexplained and the subject of intense research.

(Graphs (b) and (d) are courtesy of Shan Sun from NASA/GISS and Alex Megann from the Southampton Oceanography Centre.)

The point just made is illustrated in Figure 6, in which an observed temperature time series from the equatorial Pacific highlighting ENSO variability is compared with corresponding time series obtained from three climate models that have the oceanic component HYCOM in com-

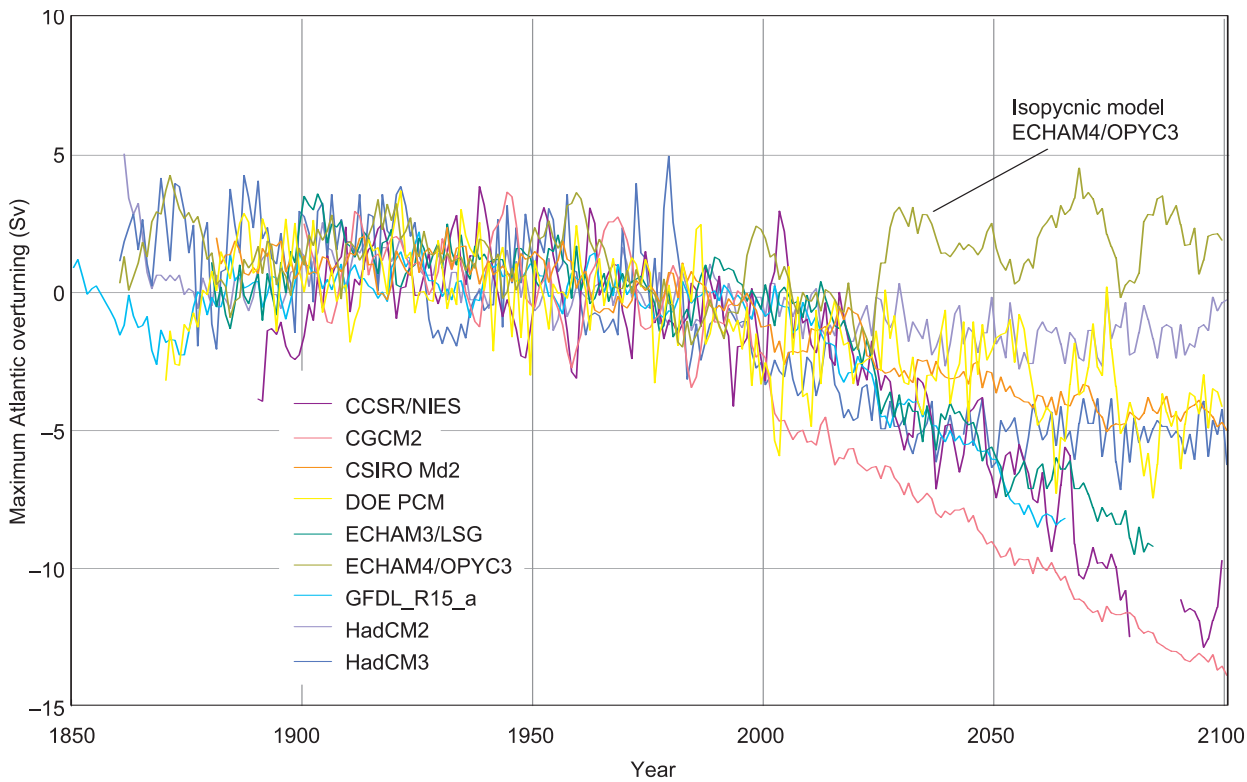


Figure 7. Effects of Global Warming in the Atlantic Overturning Rate

The curves in this plot represent changes in the Atlantic overturning rate (1 Sv = 106 $m^3 s^{-1}$) from the gradual doubling of atmospheric CO_2 in nine coupled climate models. Overturning rates are plotted relative to each model's average over the period from 1960 to 1990. A reduction by 15 to 20 Sv amounts to a total shutdown of the overturning. Whereas the eight level models show a decreasing overturning rate, the isopycnic, or layer, model ECHAM4/OPYC3 does not indicate a slowdown of that rate under the conditions described above. (Reproduced courtesy of IPCC 2001.)

mon. As expected, the amplitude of the ENSO mode depends on which atmospheric-model component the ocean is coupled to. But the large difference in the two GISS/HYCOM results (b) is mainly caused by changing the target densities and vertical mesh spacing in HYCOM. A grid configuration that minimizes the vertical extent of the depth-coordinate subdomain in the eastern Pacific, thereby allowing the isopycnic subdomain to rise close to the surface, seems to favor large-amplitude El Niño variability in the model. It is tempting to attribute this phenomenon once again to the superior thermal insulation properties of the isopycnic vertical coordinate.

As stressed in the introduction, a model may well be able to satisfactorily predict long-term global change caused by extraneous factors such as increased greenhouse gas concentrations even if it does a less-than-perfect job in simulating ENSO.

Atlantic Overturning during Global Warming. Changes in ocean circulation, particularly in the strength of the meridional overturning circulation (MOC) in individual basins, are considered plausible triggers of rapid climate change (Broecker 2003). What began as a highly technical discussion of this issue has recently seeped into more popular publications (*Fortune*, February 26, 2004; *The Observer*, February 22, 2004). Given the pivotal role played by the Atlantic in moving heat to high northern latitudes (as highlighted in Figure 2), climate researchers are keenly interested in processes that have led to a periodic weakening or outright shutdown of the Atlantic MOC since the last ice age. Foremost among the processes that can trigger such effects is the buildup of a freshwater cap in the subpolar Atlantic by melting land and sea ice. Since seawater density at near-freezing temperatures depends almost entirely on salinity, accelerated

ice melt during global warming could conceivably create a strong enough vertical density contrast in the subpolar Atlantic to inhibit the sinking of surface water to the bottom, thereby suppressing the MOC.

Such a shutdown can easily be simulated in an ocean model by imposing an appropriate high-latitude freshwater source. The question is, "how robust a feature is the Atlantic MOC in a climate model?" In other words, is the threshold for an MOC shutdown by ice melt in the model the same as the threshold in the real ocean? Figure 7, taken from the 2001 climate assessment report by the Intergovernmental Panel on Climate Change, indicates that there are vast differences among models in predicting the rate at which the Atlantic MOC will slow down during global warming. Interestingly, from among nine climate models, only an isopycnic coordinate, or layer, model does not indicate a slowdown of the MOC during gradual doubling of atmospheric carbon dioxide (CO_2). This observation suggests that the type of vertical coordinate in an ocean model can greatly influence the outcome of a climate forecast—for reasons touched upon earlier in this article. Further support for the still tentative notion that layer models predict a more stable behavior of the Atlantic MOC during global warming than the eight level models shown in Figure 7 can be found in Sun and Bleck (2001b). Note, however, that the jury is still out on whether "more stable" is synonymous with "more realistic."

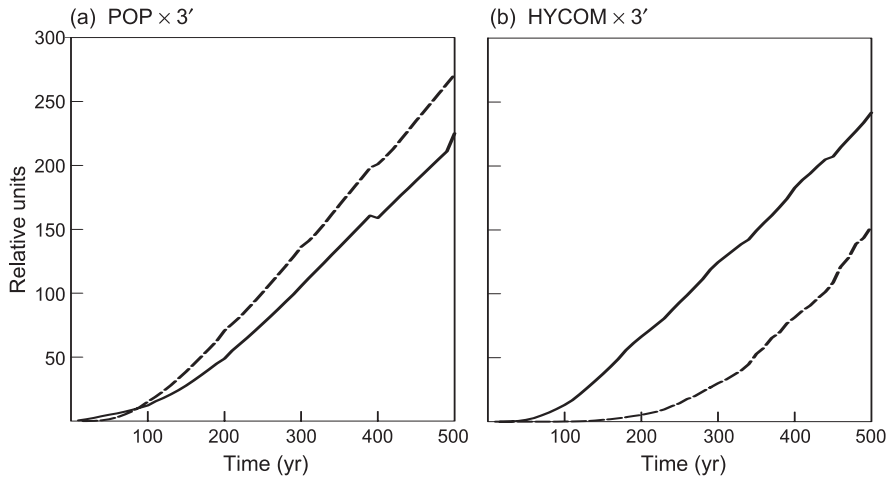


Figure 8. Comparing Carbon Sequestration Results
The two plots show the gradual accumulation of tracer material representing CO_2 (arbitrary mass units) in the top 10 m of the world ocean in (a) POP and (b) HYCOM. The conditions were continuous tracer release at two near-bottom points next to the North American continental shelf off Delaware and California. The tracer injected off Delaware is represented by the solid line; the one injected off California by the dashed line. The source strength at both sites is 1 mass unit per day (36,000 units per century). The discrepancies between the POP and HYCOM results are a manifestation of the uncertainty attributable to numerical approximations in ocean circulation models.

Transport of Sequestered CO_2 in the Ocean. A standard question asked of a climate prediction model is whether its "equilibrium" climate, obtained by running the model for a long time (several centuries) with a time-invariant mixture of atmospheric greenhouse gases and constant solar-energy output (that is, with fixed boundary conditions), resembles the observed climate. Given a century or two, an energy imbalance of a few watts per meter squared, less than 1 percent of the standard solar energy input, will gradually melt the polar ice caps or bring on an ice age in the model. Since the heat capacity of the atmosphere is negligible compared with that of the ocean, radiative imbalances are primarily accumulated in the ocean (including its frozen component). There they create long-term trends in the thermal structure, which sooner or later will disrupt the overturning circulation and the associated poleward heat transport. Interestingly, the drift in global surface temperature accompanying these changes may be as small as a fraction of a degree. (That is why sea surface temperature maps, often presented as an indicator of the performance of an ocean model, are of limited usefulness.)

Much time is being invested at the Laboratory and elsewhere into studying the sensitivity of the modeled MOC to changes in the boundary conditions ("forcing") at the sea surface. In many of these studies, for the sake of computational economy and to avoid contaminating the ocean simulation with atmospheric model errors, the ocean is driven by observed values of temperature, precipitation, wind, or other factors rather than by an atmospheric model that properly reacts to the evolving surface conditions in the ocean model. However, replacing an interacting atmospheric model with prescribed surface fields elicits unforeseen responses in the ocean model. Efforts at Los Alamos to compare the performance of layer and level models in ocean-only experiments have been frustrated by the realization that ocean models show different degrees of tolerance to physically imperfect surface forcing.

Nevertheless, enough progress has been made over the years in formulat-

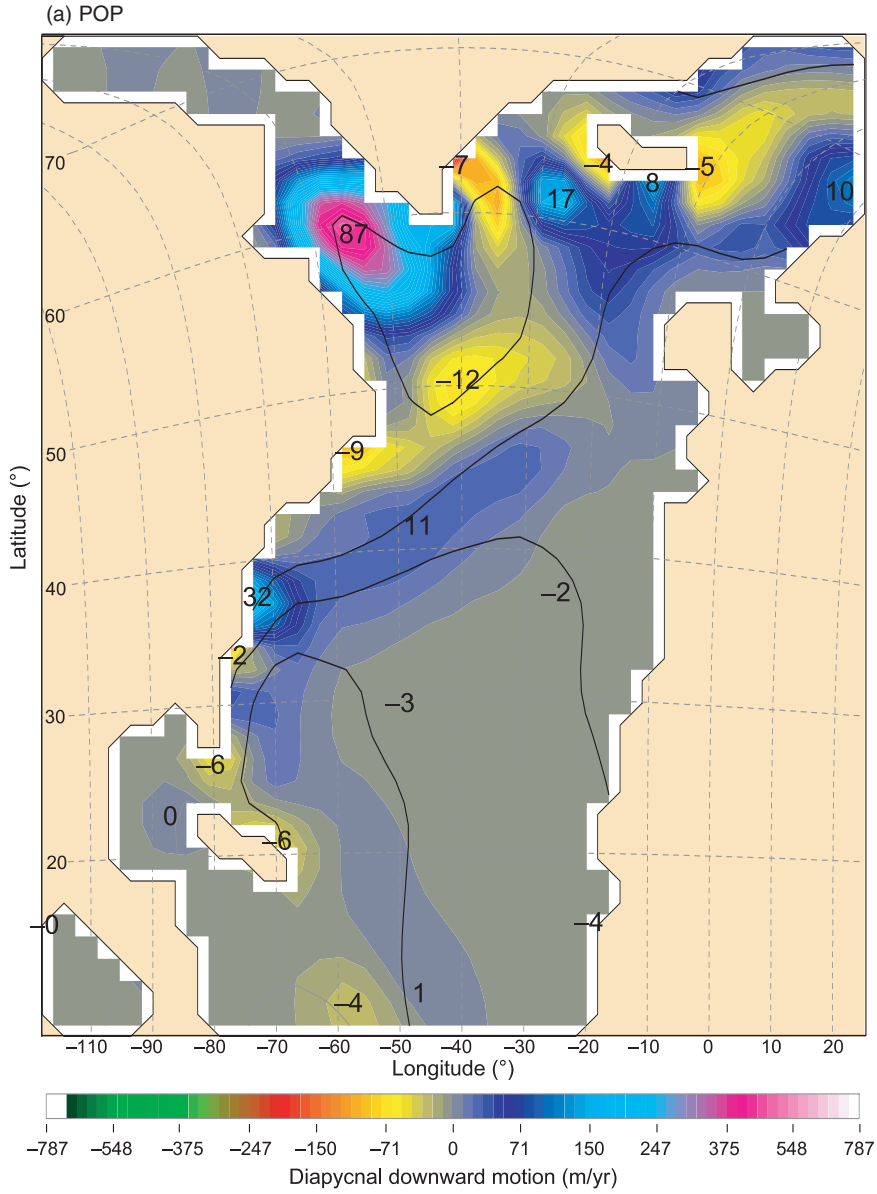


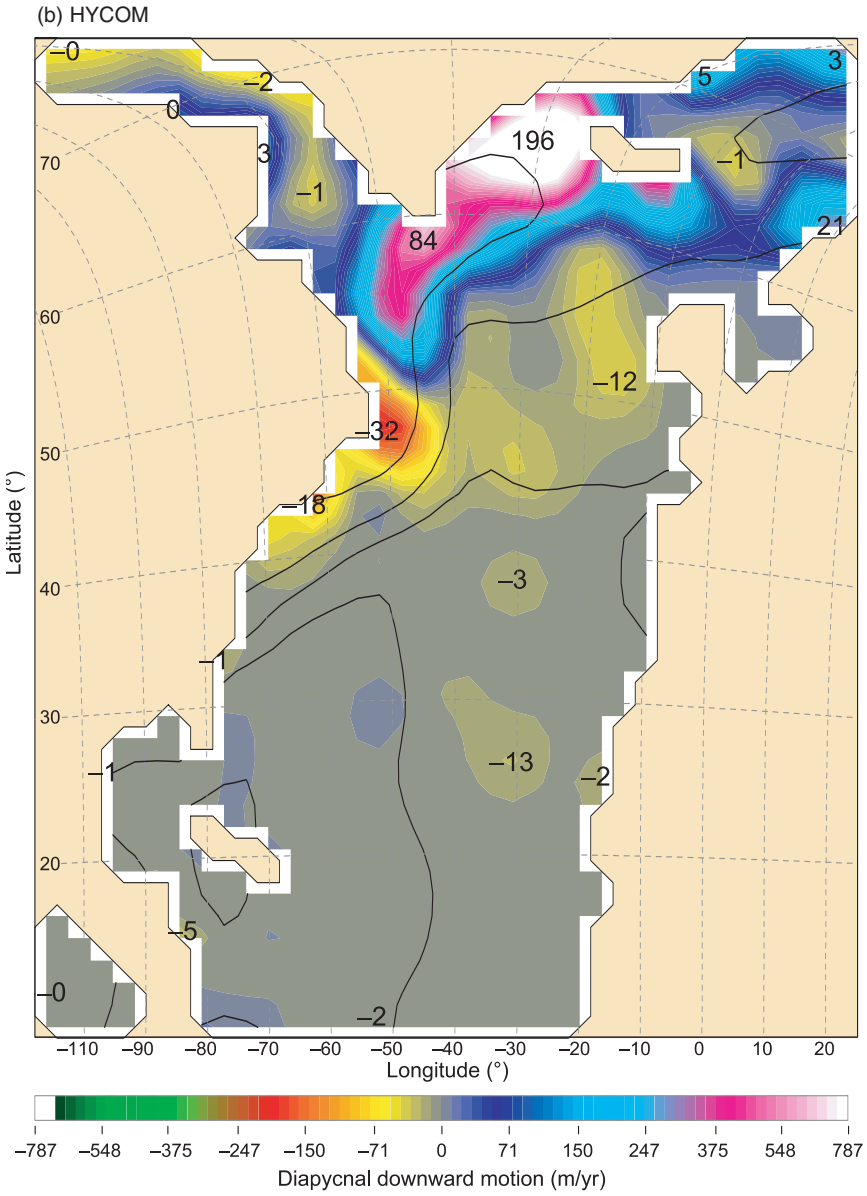
Figure 9. The Downwelling Limb of the Atlantic Overturning Circulation in POP and HYCOM
These isopycnic-coordinate views of the thermally forced Atlantic circulation were obtained with two coarse-mesh models: (a) POP and (b) HYCOM. Each view shows North America at left and Europe and Africa at right. Greenland (grossly deformed by the map projection) is seen at the top. Color contours represent the time-averaged rate (meters per year, positive downward) at which water crosses an isopycnic surface near the interface between the warm and cold limbs of the Atlantic overturning circulation. Numbers overlaying the patches of upwelling and downwelling indicate the total diapycnal mass flux (in units of 0.1 Sv, positive down) associated with each patch. Also shown are sea-surface height contours (at 20-cm intervals), a proxy for streamlines of surface currents. The figure illustrates that numerically dissimilar ocean models will disagree on the strength and geographic distribution of the downwelling limb of the overturning circulation even when subjected to identical surface boundary conditions.

ing internally consistent surface boundary conditions to produce reasonably steady and realistic equilibrium circulation states in ocean-only experiments. These circulation states can be used for a variety of practical applications, among them studies of the efficacy of CO_2 sequestration in the world ocean.

Among options currently under discussion for slowing down greenhouse gas-induced global warming is pumping liquefied CO_2 into the abyssal ocean. Regardless of the potential ecologic side effects or the economic feasibility of this approach—not to be discussed in this article—oceanic carbon sequestration presents an interesting test case for studies aimed at comparing ocean models.

One question that can be addressed through numerical simulation is how much time it would take for CO_2 injected into the deep ocean to come back to the surface. Figure 8 shows the results of such a simulation in which an inert tracer representing CO_2 is continually being released close to the sea floor at two points located at 37°N near the continental shelf off the American East and West Coasts. The curves show the globally averaged near-surface buildup of that tracer as it gradually works its way through the global ocean. This buildup provides a semi-quantitative measure of how soon the sequestered CO_2 is likely to re-enter the atmosphere through transport and diffusion alone.

The experimental details can only be sketched here. The simulation is performed with both POP and HYCOM configured on the same



noneddy-resolving horizontal grid and subjected to identical seasonally varying atmospheric forcing. POP uses 25 levels in the vertical direction, whereas HYCOM uses 16 layers. The tracer is transported "offline" using the two models' seasonally varying circulation states averaged over consecutive 3-month intervals. The offline approach is chosen for computational economy. The time step in most fluid models is set by the time it takes for the fastest signal supported by the model equations to propagate from one grid point to the next. In the ocean model, the fastest signals (gravity waves) travel in excess of 200 meters per second (ms^{-1}), but advection by currents is at least 100 times slower. Hence, offline tracer advection, in which gravity waves are not an issue, can be done with a 100 times longer time step, and hence 100 times faster than in the full ocean model itself.[2]

In preparation for tracer transport, horizontal mass fluxes from both models are transformed into isopycnic fluxes (fluxes along isopycnic surfaces) from which the missing diapycnal component is deduced by mass continuity. This transformation is performed to ensure that global ocean-ventilation processes, whose action is modeled most coherently in isopycnic coordinates, act similarly in both models. Plots of diapycnal mass flux fields (Figure 9) indeed indicate that both models maintain an Atlantic overturning circulation that is in fair agreement with the available observational evidence (refer to Figure 2).

The vertical flux fields in Figure 9 are a study in model-to-model variability in their own right. Both models clearly depict the ocean basins surrounding southern Greenland as the region anchoring the downwelling limb of the Atlantic overturning circulation, but differences in local detail are obvious. Note that vertical motion is analyzed here in potential-density space; hence, it depicts areas where individual seawater parcels get either lighter or denser with time. Consequently, upwelling and downwelling patches in Figure 9 coincide with regions where the ocean exchanges heat with the atmosphere. Given that atmospheric cyclones thrive on surface heating (the notorious Cape Hatteras storms are a good example), the different MOC downwelling patterns indicated in Figure 9 are likely to result in large differences in regional weather. Storminess in the Irminger Sea, east of Greenland, for example, would be affected by the surface heat-flux differences indicated in Figure 9(a).

[2] Ongoing efforts at Los Alamos and elsewhere try to lengthen the time step in ocean models by filtering out gravity waves, but the ensuing mathematical complexities are daunting. Gravity waves do serve a purpose, both in reality and in the model: They repair deviations from "geostrophic" equilibrium, a particular balance between velocity and pressure field, on which fluids on a rotating planet rely to counteract the deflecting effect of the Coriolis force.

Overall, the Atlantic MOC appears to be stronger in HYCOM than in POP, consistent with the earlier discussion about differences in vertical diffusion control in level and layer models. Reduced momentum mixing in the vertical direction, that is, lower drag on the wind-driven surface flow, may also be the cause for the somewhat stronger surface circulation in HYCOM. This difference is indicated in Figure 9 by the tighter spacing of sea-surface height contours in the right panel compared with those in the left panel. In geostrophically balanced flow, sea-surface height contours are a proxy for streamlines, like isobars on a weather map.

The salient result from Figure 8 is that, after 500 years, POP has brought 1.5 times more material sequestered off California back to the surface than HYCOM. Model-to-model differences are much smaller for the material sequestered off Delaware. Since the circulation off the U.S. East Coast is dominated by strong opposing boundary currents representing the cold and warm limbs of the Atlantic MOC, a feature not found off the West Coast, large differences in dispersion from the Pacific and Atlantic release sites are to be expected. HYCOM accentuates those differences more than POP.

These results, which represent ongoing work and remain to be confirmed by additional experiments, are tendered here as a first attempt at quantifying the circulation-related uncertainties in simulating the feasibility of abyssal sequestration of CO_2. These uncertainties are compounded, of course, by uncertainties about the chemical behavior of CO_2 at great depths.

Concluding Remarks

This article has presented some of the tools used by the research community to assess the uncertainty in decadal to century-scale climate prediction. For the discussion, climate prediction has been cast as a boundary-value problem in which the boundary values of interest (forcings) are assumed to be known. In other words, forcing uncertainties, which are a major point of debate in their own right, have not been considered. Instead, the focus in this article is on error sources within climate models. Limiting the number of climate-relevant natural processes, as well as parameterizing processes that are deemed important but take place on scales too small to be resolved by the model's space-time mesh, creates one type of errors: type 1, or physical-closure, errors. The conversion of the underlying differential equations into computer-solvable algebraic equations, which mainly affect processes the model is designed to resolve explicitly, results in another type of errors: type 2, or numerical, errors.

To guide future model development, the effects of these two error types on the performance of a climate model need to be separated. At least in principle, one can separate those effects either by manipulating type-1 errors (by, for example, adding/subtracting earth system processes or refining certain physical closure schemes) or by quantifying type-2 errors through solving the same physical problem with numerically dissimilar models. Unfortunately, experimenting with different mesh sizes in a climate model—the approach usually taken to establish the proximity of a numerical to a "true" solution—typically does little to disentangle the two error types because physical closure assumptions often are tailored to a particular mesh size and are not expected a priori to work well if the resolution is changed.

Los Alamos is making important contributions in this area by supporting the development and use of multiple ocean models in climate simulation. The numerical diversity in the Laboratory's model ensemble is achieved by support of both level and layer ocean models. The former discretize the underlying differential equations on a Cartesian grid whereas the latter use a material, or Lagrangian, vertical coordinate tied to the oceanic potential-density field, a proxy for entropy. Vertical dispersion of physical properties is handled very differently in these two types of models. Because subsurface oceanic processes are adiabatic (except for mixing) and hence are governed by the entropy conservation law on centennial time scales and beyond, numerically different approaches to satisfying the second law of thermodynamics can lead to profoundly different equilibrium circulation states in long-term ocean simulations. In fact, the sensitivity of the model solution to discretization (type-2) errors in the thermodynamic and dynamic equations often overshadows the sensitivity to the physical-closure assumptions (type-1 errors).

Shortcomings of layer models having to do with the difficulty of defining constant-density surfaces in unstratified regions (regions in which water density does not vary with depth) have led to the development of so-called hybrid-coordinate models, which also are included in the ocean model mix at Los Alamos. ■

Acknowledgments

The climate modeling and carbon sequestration work presented here is being funded by the Climate Change Prediction and Ocean Science Programs, respectively, of DOE's Office of Science, Climate Change

Research Division. The author is indebted to Shan Sun (NASA/GISS) and Alex Megann (Southampton Oceanographic Centre) for providing the Niño3 index diagnostics shown in Figure 6, and to Mathew Maltrud (Los Alamos National Laboratory, Theoretical Division) for providing POP circulation data used in generating the results shown in Figures 7 and 8. Careful scrutiny of this article by members of the COSIM team has led to significant improvements in the presentation.

Further Reading

AchutaRao, K., and K. R. Sperber. 2002. Simulation of the Elf Niño Southern Oscillation: Results from the Coupled Model Intercomparison Project. *Clim. Dyn.* **19** (3–4): 191.

Bleck, R., C. Rooth, D. Hu, and L. T. Smith. 1992. Salinity-Driven Thermocline Transients in a Wind- and Thermohaline-Forced Isopycnic Coordinate Model of the North Atlantic. *J. Phys. Oceanogr.* **22**: 1486.

Bleck, R. 2002. An Oceanic General Circulation Model Framed in Hybrid Isopycnic-Cartesian Coordinates. *Ocean Model.* **4** (1): 55.

Broecker, W. S. 1991. The Great Ocean Conveyor. *Oceanogr.* **4**: 79.

Hirt, C. W., A. A. Amsden, and J. L. Cook. 1974. An Arbitrary Langrangian-Eulerian Computing Method for All Flow Speeds. *J. Comput. Phys.* **14** (3): 227.

IPCC. 2001. Climate Change 2001: The Scientific Basis—Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Edited by J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, et al. Cambridge: Cambridge University Press.

Malone, R. C., R. D. Smith, M. E Maltrud, and M. W. Hecht. 2003. Eddy-Resolving Ocean Modeling. *Los Alamos Science* **28**: 223.

Schmitz Jr., W. J. 1996. "On the World Ocean Circulation. Vol. 1, Some Global Features/North Atlantic Circulation." Woods Hole Oceanogr. Inst. Tech. Rep. WHOI-96-03.

Sun, S., and R. Bleck. 2001a. Thermohaline Circulation Studies with an Isopycnic Coordinate Ocean Model. *J. Phys. Oceanogr.* **31** (9): 2761.

———. 2001b. Atlantic Thermohaline Circulation and Its Response to Increasing $C0_2$ in a Coupled Atmospheric—Ocean Model. *Geophys. Res. Lett.* **28** (22): 4223.

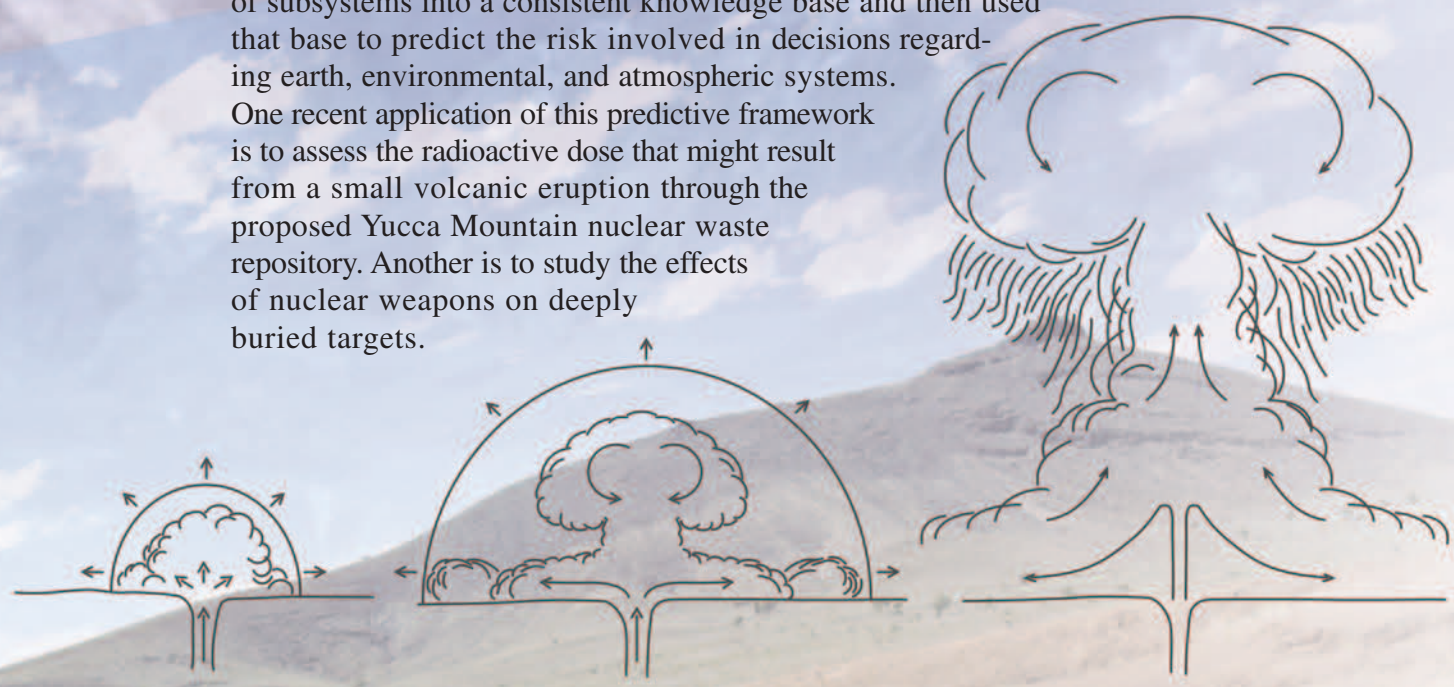
For further information, contact Rainer Bleck (505) 665-9150 (bleck@lanl.gov).

Predicting Risks in the Earth Sciences

Volcanological Examples

Greg Valentine

Where can nuclear waste be safely placed? How can humans better manage natural resources? How can humans prevent manmade disasters and prepare for natural ones? Sound decisions require knowledge of the subsystems in each problem and a reliable decision-making framework. Over the last several decades, earth scientists at Los Alamos have integrated experiment, observation, and modeling of subsystems into a consistent knowledge base and then used that base to predict the risk involved in decisions regarding earth, environmental, and atmospheric systems. One recent application of this predictive framework is to assess the radioactive dose that might result from a small volcanic eruption through the proposed Yucca Mountain nuclear waste repository. Another is to study the effects of nuclear weapons on deeply buried targets.



(Left to right) First three photos are courtesy of J. Hughes, J. Franklin, and R. McGimsey, respectively. The last photo is courtesy of the U.S. Geological Survey.

Prediction is at the heart of applying earth science to issues of importance to society. A common application of predictive earth sciences is weather forecasting, which is particularly important to mitigating the consequences of severe weather. Other applications include global climate change, availability and quantities of natural resources, natural disaster planning and mitigation, performance of geologic repositories, and nuclear weapon effects. Each of these applications involves systems that are composed of many subsystems; for example, global climate change depends on cloud physics, mass and energy transport between the biosphere and atmosphere, ocean dynamics, and anthropogenic processes, to name only a few. These subsystems may be coupled to each other through nonlinear processes and across a wide range of time and space scales. Data on the subsystems are collected at varying resolutions, and none of the subsystems is fully characterized; in addition, many of the predictions we are interested in often involve extreme rather than normal conditions for the systems or subsystems. All these aspects contribute to an inherent uncertainty in predictions. Finally, the only information we have on the behavior of fully coupled systems, such as climate, is historical; we cannot do controlled experiments on the full systems. Significantly, all the features mentioned above, namely, nonlinearly coupled subsystems, multiple scales, uncertainty, extreme conditions, and an inability to experiment on full systems (except for analyzing

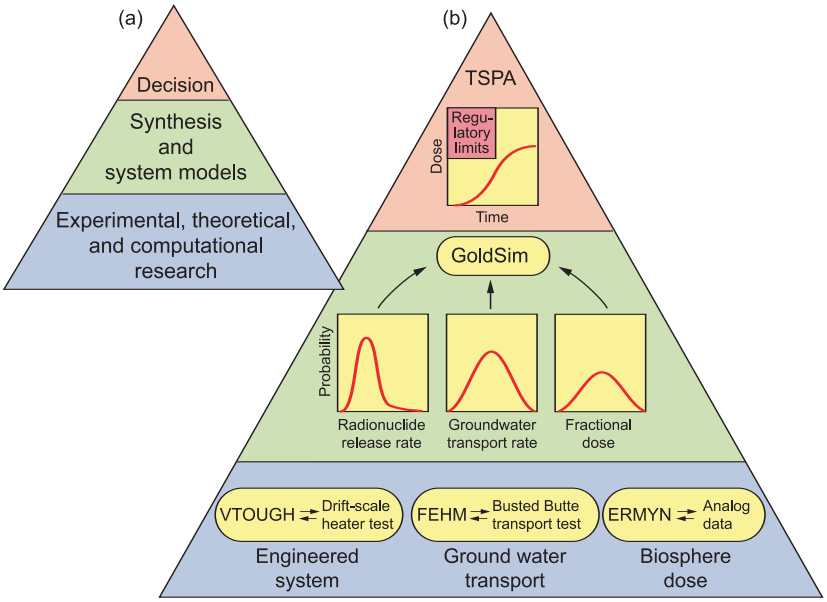


Figure 1. Framework for Predictive Earth Sciences
(a) The framework for predictive earth sciences illustrates the foundation in fundamental experimental, observational, and theoretical and/or computational research on the basis of which decisions are made. (b) Illustrated at right is a specific example for predicting dose from potentially contaminated ground water at Yucca Mountain, showing some components of the multiple-barrier repository system that have been studied in detail by combined experimental and theoretical approaches. For example, the engineered part of the system includes, among other things, the walls of tunnels (or drifts) in the mountain that will experience heating (due to the radioactive decay of the waste) and resulting mass transfer processes. These have been studied with the VTOUGH code coupled with observations from a full-scale test (drift-scale heater test), whereby mock waste packages were emplaced in a tunnel and heated, while the temperature and mass transport were monitored in the tunnel walls. The next barriers that leaking radionuclides would encounter is the thick zone of unsaturated (pore spaces are not completely filled with water) rocks above the water table and then by the saturated zone below the water table, which provides a pathway to a hypothetical future population some 18 km away. Tests such as the Busted Butte transport test, in which surrogates for radionuclides were injected into unsaturated rocks and their migration was monitored, are coupled with codes (for example, the Los Alamos FEHM code) that simulate the detailed physics of flow and transport through rocks. Finally, studies have been conducted to determine the potential radioactive dose a human might receive from any radionuclides that might have migrated sufficiently far. Those studies combined the dose code ERMYN with analog information (for example, studies of dose from atmospheric nuclear testing fallout). The results and uncertainties of these subsystem studies and detailed predictions are then abstracted and integrated with a simulation package (Goldsim) produced by the GoldSim Technology Group, LLC, to produce a prediction of dose as a function of time.

historical data) are similar to the core features that make predicting the reliability of our nuclear weapons stockpile a challenging process (Valentine 2003).

Predictive earth sciences involve the integration of experiment, observation, and modeling to form the basis for decisions involving earth, environmental, and atmospheric systems. Figure 1(a) illustrates the main elements of predictive earth sciences in the form of a pyramid. The foundation for predictions is built upon fundamental experimental (including observations), theoretical, and computational research into the behavior of individual subsystems and, as appropriate, the coupling between them. For some subsystems, the necessary information can be obtained from experimental data, but most of the complex subsystems that we work with involve an iterative approach among experiment, observation, theory, and computation. Once we have an adequate understanding of the important subsystems, we synthesize and simplify that information, accounting for uncertainties, and build it into a system model. The system model accounts for all the couplings between subsystems and their uncertainties, and produces a probabilistic prediction of system behavior that can be used for decision-making.

Figure 1(b) illustrates this framework with a specific example from predicting the performance of a high-level radioactive waste repository at Yucca Mountain, Nevada. Nuclear Regulatory Commission regulations define repository performance in terms of radiation dose to a human population at a location 18 kilometers south of the repository over a period of 10,000 years. In the absence of an unusual, disruptive event, a dose can be received only if radionuclides escape through a series of engineered and natural barriers. Among the engineered barriers are glass or ceramic pellets embedded with radioactive spent fuel, cladding that covers the waste, and storage canisters containing spent fuel rods laden with highly radioactive fission products. Water may eventually seep through the repository, corrode the canisters or cladding, dissolve the radionuclides, and carry them into the surrounding rocks. At that point, Mother Nature will have to help contain the waste. Three key natural features make Yucca Mountain desirable as a burial site for nuclear waste: its dry climate, deep water table, and thick water-unsaturated rocks above the water table. The first minimizes water that could seep through the repository and eventually corrode the waste canisters. The second enables building a repository that is deep underground (300 meters) yet still well above the water table, which is another 240 to 300 meters lower. The third natural feature is a thick zone (several hundred meters thick) of water-unsaturated rocks containing clays, zeolites, and other minerals that adsorb numerous radionuclides and thus effectively slow down leakage of radionuclides into the water table.

If, in spite of these features, radionuclides were to be transported by ground water to the control population, the contaminated water might then be pumped and used for drinking or irrigation of crops, which are pathways for human dose. Within the predictive-earth-sciences framework, each of these barriers or steps in the movement of radionuclides is a subsystem, some of which are shown in Figure 1.

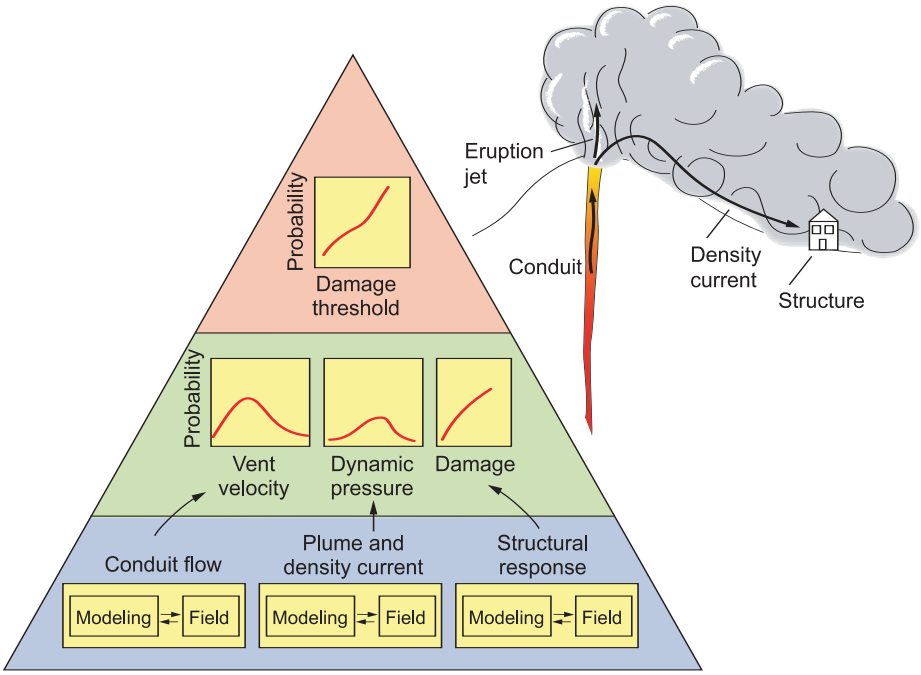


Figure 2. Predicting Volcanic Risk to Buildings
The predictive earth-science framework shows some of the important components used for predicting risk to buildings from explosive volcanic eruptions. Among these components are the dynamics of flow up the volcanic conduit (red), which determines the initial and boundary conditions for an eruption, the rise and collapse of an eruption jet or volcanic plume and the resulting pyroclastic density current (gray), and the response of building structures to the conditions produced by the currents. Predictions of these individual components, combined with the probability of a volcanic event and with other components that are not discussed here, must ultimately be integrated to produce a probabilistic prediction of damage thresholds that might be exceeded in developed areas around a volcano.

Each of these subsystems has been studied through a closely integrated series of experiments and/or analog observations and through numerical modeling. For example, processes associated with coupled heat (from radioactive decay), which occur in the engineered part of the system fluid flow, in porous and fractured rocks, and in reactive chemical transport within those fluids, have been approached with an experimental program known as the Drift-Scale Heater Test and with the computer code VTOUGH. The test is a full-scale mockup of a heated waste package placed in a tunnel, where instruments measure mass and energy fluxes in the surrounding rocks; the computer code was written at Lawrence Berkeley National Laboratory and was modified by researchers at Lawrence Livermore National Laboratory to simulate the engineered barrier system.

Ground-water flow and radionuclide transport within the unsaturated zone beneath the repository have been studied from results of field-scale experiments such as the Busted Butte transport test and with the finite element heat and mass (FEHM) transport code (Eckhardt et al. 2000). The latter has also been used to study the saturated zone. Actual conversion of the transported radionuclides into human dose has been constrained with analog data and the ERMYN code (BSC 2004). In the simplest sense, the predictions of each of these subsystems are cast into probability distributions of the parameters of interest—for example, the rate of radionuclides released from the engineered system, the rate of radionuclide transport by ground water to the human population, and the fraction of radionuclides from that ground water that is taken in by humans as dose. The probabilistic approach allows us to incorporate the uncertainties inherent in each subsystem. These distributions are then sampled with a Monte Carlo software engine (for example, GoldSim, which was developed by the GoldSim Technology Group) to produce a simple plot of dose to humans as a function of time, as shown at the top of the diagram. If the predicted dose (which might be the mean value of a large number of realizations, representing uncertainties) crosses over the regulatory limit (represented by the yellow box), the repository is not feasible. Thus, a large amount of complex science on the behavior of numerous subsystems is boiled down into a simple answer, which is directly used by decision makers. The framework shown in Figure 1 is iterative between the apex and the base—in other words, the framework can be reversed to decide which subsystems produce the greatest sensitivity in the final result and therefore might need further research to reduce uncertainties.

The predictive-earth-sciences framework is also being applied to assessing risk from explosive volcanic eruptions. The main body of this article will cover a few of the important components of the volcanic risk problem (see Figure 2). Ultimately, risk is determined by the probability of an event occurring, combined with the probability of damaging effects on humans, buildings, or other infrastructure (Perry et al. 2000; Valentine 1998 and 2003). A chain of events, or subsystems, determines the damaging effects, such as flow of magma up a conduit in the earth's crust, eruption into the air as a jet of gas and particles or clots of magma, and subsequent flow of that mixture across the landscape as a density current. The next few sections will describe models of the three subsystems. Although they are work in progress, our models demonstrate the synergy that must exist among theory, experiment, observation, and computation when predicting complex systems. The last section will also show results of an integrated volcanic-risk assessment that follows the predictive-earth-science framework but with simpler subsystem models than the ones referred to above. This assessment combines both probability of occurrence and the consequences of a potential volcanic event at the proposed Yucca Mountain, Nevada, high-level radioactive waste repository. Finally, the article will discuss how the predictive-earth-science framework can be applied to other problems of importance for both military application of nuclear weapons and energy security.

Conduit Flow Models and Quantification through Field Studies

Eruption processes are determined by the velocity, pressure, temperature, and gas content of material exiting a volcanic vent; these, in turn, are determined by processes in the subsurface. At some depth beneath a volcano (typically between 5 and 30 kilometers), magma accumulates in what is typically referred to as a magma chamber. The magma, which is a mixture primarily of silicate melt, crystals, and bubbles, will contain several dissolved gases, or volatiles, of which water (H_2O) is the most abundant in most cases. As magma rises through a conduit toward the earth's surface, it experiences successively lower pressures with decreasing rock overburden. Because the solubility of volatiles in the magma decreases with decreasing pressure, volatiles that were dissolved at magma chamber depth will come out of solution to form bubbles of gas. As the magma continues to rise and decompress, it releases more volatiles into bubbles, and the bubbles expand. In order to conserve mass, the expanding mixture must accelerate. This acceleration is also determined by the conduit dimensions. The expansion of the magma

mixture and the conduit dimensions are ultimately coupled because the walls of the conduit might be eroded by the magma as it accelerates.

Using a multifield approach for modeling the upward flow of magma, whereby gas and melt are treated as overlapping continua that are coupled by mass, momentum, and energy exchange, Macedonio et al. (1994) developed a system of governing equations to describe conduit flow, the first component for predicting volcanic risk illustrated in Figure 2. The equations (see box at right) include several simplifying assumptions: one-dimensional, steady flow; constant conduit geometry (which assumes that wall-rock materials introduced into the flow are not in sufficient quantities to change the shape of the conduit appreciably); and isothermal flow (thus the lack of an energy conservation equation). However, the equations do account for the rise of separate gas and droplet/particle (incompressible) phases, frictional coupling between those phases, and the introduction of wall-rock debris into the mass and momentum balances. The term C_w, the mass erosion rate of wall rock per meter into the flow, accounts for the interaction between the flow and the conduit walls. Because the flow is considered to be one-dimensional, steady, and in a constant-geometry conduit, it is implied that the mass erosion rate is small. In reality, there might be more erosion that sufficiently changes the conduit shape to negate the simplifying assumptions in these equations. The current treatment should be regarded only as a first step toward addressing the difficult problem of fully coupled flow and solid walls.

Given the wide range of conditions within volcanic conduits and the even wider range of potential wall-rock properties, C_w is difficult to constrain theoretically. For that reason, we designed a series of field studies to provide quantitative values for C_w at extinct volcanoes in the southwestern United States. Field sites were selected according to criteria that allow quantification of the amount of wall-rock debris as a function of depth below the volcanoes: (1) The volcanoes must be old enough that many of their deposits are exposed by erosion, or the deposits might be exposed by quarry operations; (2) the sequence of rocks below the volcanoes must be well constrained in terms of the thickness of individual layers; (3) fragments of those layers should be easily identifiable in the volcanic deposits recording the eruptions; and (4) the different styles of eruption processes must be easily interpreted from the volcanic deposits. At sites that meet these criteria, it is then possible to measure the volume fraction of fragments from each layer of wall rock within volcanic deposits; dividing that value by the thickness of the layer results in an average volume fraction per meter.

Several volcanoes in two regions, the Lucero Volcanic Field of west-central New Mexico and the San Francisco Volcanic Field of northern Arizona, meet these criteria. Wall-rock erosion data from the Lucero Volcanic Field, in particular, illustrate the variations in wall-rock erosion for eruption mechanisms that range from relatively passive eruption of lava to Hawaiian-style lava fountains and from those to very violent eruptions involving explosive interaction of magma (at about 1100°C) with ground water. Figure 3 shows the volume fraction per meter (erosion rate) for the latter type of eruption (left side) and for more passive types (right side), corresponding to the layers of wall rocks beneath the volcanoes. Erosion rates vary over factors of 1000 to 10,000, depending upon the

Conduit Flow Model

Equations of Macedonio et al. (1994)

Mass

$$G_G = \rho_G \alpha u_G \quad \text{Mass flow gas per unit area}$$

$$\frac{dG_L}{dz} = C_w \quad \text{Change in particle mass flux/area due to wall-rock erosion } (C_w = \text{mass erosion rate})$$

Momentum

Gas

$$\rho_G u_G \alpha \frac{du_G}{dz} = -\alpha \underbrace{\frac{dP}{dz}}_{\text{Pressure gradient}} - \underbrace{F_{LG}}_{\text{Friction with particles}} - \underbrace{F_{WG}}_{\text{Friction with walls}} - \underbrace{\rho_G g \alpha}_{\text{Gravity}}$$

Particles

$$\rho_L u_L (1-\alpha) \frac{du_L}{dz} = -(1-\alpha) \underbrace{\frac{dP}{dz}}_{\text{Pressure gradient}} + \underbrace{F_{LG}}_{\text{Friction with particles}} + \underbrace{F_{WL}}_{\text{Friction with walls}}$$

$$- \underbrace{\rho_L g (1-\alpha)}_{\text{Gravity}} + C_w (u_w - u_L)$$

Assumptions: Steady state, 1-D, isothermal

eruption mechanism and the types of wall rock. These rates can be used to constrain C_w for the conduit fluid model equations. For more details on these field studies, refer to Valentine and Groves (1996). The main point here is to show that combining theoretical and/or computational modeling with field studies will yield quantitative estimates for volcanic conduit flow, one component of volcanic risk prediction. More data are being collected and implementation of the field-derived C_w values into the numerical solution of the conduit fluid model is a future goal.

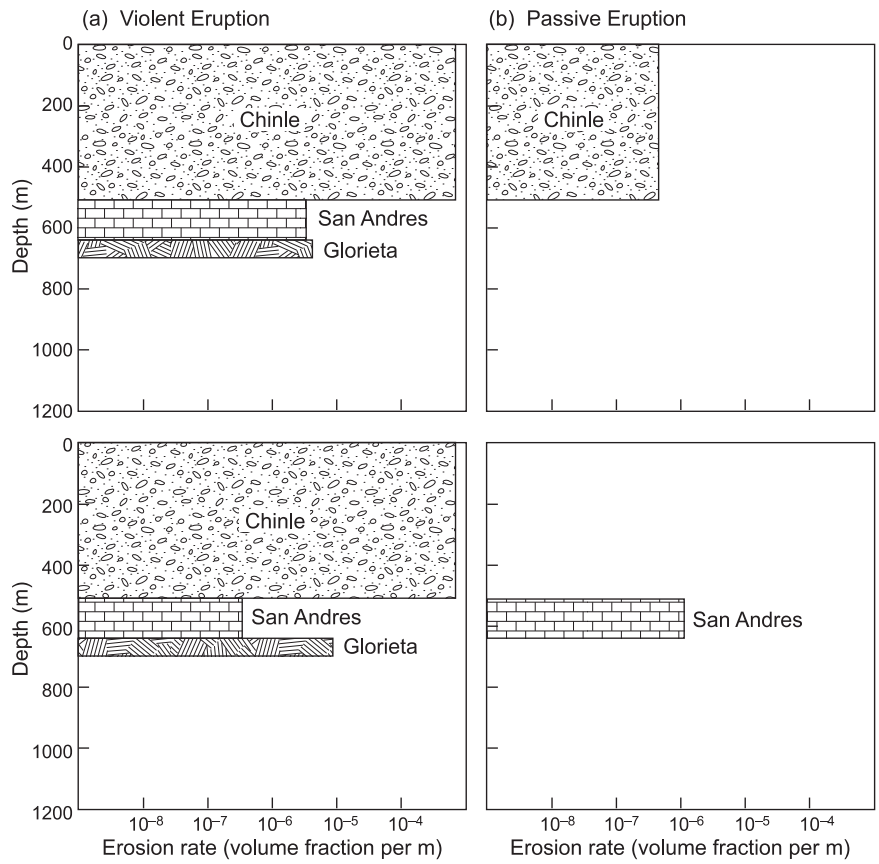


Figure 3. Data on Wall Rock Erosion from Violent and Passive Eruptions
These data on wall rock erosion are from field measurements and are expressed in terms of volume fraction per meter down the conduit for different wall-rock formations (Chinle, San Andres, and Glorieta formations) beneath volcanoes in the Lucero Volcanic Field. (The plots are adapted from Valentine and Groves 1996. Entrainment of Country Rock During Basaltic Eruptions of the Lucero Volcanic Field, New Mexico. *J. Geol.* 104 (1): 71, published by the University of Chicago.)

Plume and Density Current Models

The next process illustrated in Figure 2 is the prediction of volcanic plumes and pyroclastic density currents (PDCs) (the word "pyroclastic," from the Greek roots for fire and broken, refers to the fragments of quenched magma, such as pumice and smaller fragments misleadingly called ash, as well as fragments of wall rocks that are ejected during explosive eruptions). The volcanic plumes of interest consist of gas (mainly steam that has exsolved from the melt during conduit ascent) mixed with particles or clots of magma. The temperatures of these plumes when they exit the volcano are typically about 1000°C, but the plumes are denser than the atmosphere because particles are present. Flow speeds at the vent are a few hundred meters per second, and the flows are highly turbulent. Despite being denser than the surrounding air, the plume will rise because of its initial momentum. As it rises, it will decelerate and simultaneously mix with and heat ambient air such that the overall mixture density decreases. Sustained volcanic plumes exhibit two end members of behavior that depend upon the flow conditions as the flow exits the conduit (these conditions are calculated with a model such as the one discussed in the preceding section). In one end member, the plume is able to mix with sufficient air that, by the time it reaches the height at which its initial momentum has been lost, the plume is less dense than the surrounding atmosphere and continues to rise until it reaches a neutral buoyancy level (which might range from several kilometers to as much as 50 kilometers above the vent, depending on the eruption energy and on atmospheric conditions). The second end member occurs when the plume is still denser than the atmosphere at the time that it reaches the height determined by its initial momentum. The plume then collapses and forms a fountain of hot gas and particles, which in turn feeds density currents that flow out across the countryside. The conditions within these PDCs can be extremely damaging, particularly in heavily urbanized regions.

Based on field evidence (characteristics of deposits left behind), we can make several inferences about PDCs. (1) These mixtures of hot gas and particles can flow at a range of speeds from a few meters per second (m/s) to more than 300 m/s. This means that the flows cover a wide range of incompressible to compressible regimes in terms of the Mach number (note that the sound speed of typical

Figure 4. The Destructive Power of Pyroclastic Density Currents (PDCs)
This picture is of the landscape north of Mt. St. Helens after the devastating blast on May 18, 1980. The kinetic energy released during that volcanic event was equivalent to 7 Mt, and the thermal energy was equivalent to 24 Mt of explosive TNT equivalent. The field of view extends more than 10 km into the distance, over terrain with relief of hundreds of meters. Before the blast, the landscape was covered by a dense forest of large conifers. Their notable absence after the eruption attests to the destructive power of PDCs. The volcano in the background is Mt. Rainier.
(Photo is courtesy of J. Franklin, Mount St. Helens National Volcanic Monument photo library.)

gas-particle mixtures in PDCs can be significantly lower than in the surrounding atmosphere). (2) PDCs range in particle concentration from very dilute, essentially like sand storms (volume fractions less than 10^{-3}), to dense granular dispersions with particle volume fractions as high as approximately 0.5. At low particle concentrations, the particle and momentum transport mechanisms might be dominated by turbulence although mixture density gradients and basal traction zones can complicate the transport mechanisms. At high particle concentrations, the basal portions of the flows might have particle and momentum transport dominated by particle–particle collisions. The range of particle sizes (micrometers to meters) and densities—from about 500 to 3000 kilograms per cubic meter (kg/m^3)—combined with the depth scales of the flows, places the mixtures in a region that is somewhere between the applicability of simple, effective continuum approaches and discrete particle approaches. (3) PDCs can be variably affected by the topography over which they flow, sometimes channeling strongly into topographic lows and sometimes seeming to blanket highs and lows nearly equally. (4) Temperatures of PDCs can range up to approximately 1000°C. (5) The flows can be quite destructive (see Figure 4) and can travel more than 100 kilometers from their source volcanoes in some instances. All these factors make the prediction of PDCs very difficult and, potentially, extremely intensive computationally, depending upon the theoretical approach one takes.

Connecting PDCs to Nuclear Weapon Phenomenology. As a side note, there is a strong connection between our understanding of PDCs and nuclear weapons phenomenology. One of the founders of modern volcanology, the late Richard V. Fisher of the University of California at Santa Barbara, was assigned to Los Alamos just after World War II as a young member of the military. Later he was present at Bikini Atoll and witnessed the shallow-submarine Baker test. As the explosion column from Baker rose out of the water, a collar of water droplets mixed with steam and air collapsed back to the surface and moved outwards across the sea in a phenomenon that came to be known as the base surge. Twenty years later, Fisher, by then a professor and well-known interpreter of volcanic deposits, realized that some pyroclastic deposits around explosive volcanoes are produced by a base surge-like process as he had seen at Bikini. This connection revolutionized our understanding of volcanic processes and hazards in the 1960s. Indeed, for many years, the volcanic process was referred to as base surge or pyroclastic surge, following the nuclear weapons terminology. Recognition of a range of complications in the volcanic processes has eventually led us to the term pyroclastic density current. An interesting description of the evolution of these concepts and the nuclear weapons connection can be found in Fisher's autobiography (Fisher 1999).

Multiphase Eruption Modeling. In recent years, an important thrust in the volcanological community has been the application of multiphase flow theory to predict the behavior of eruption plumes and PDCs. This approach originated at Los Alamos in the 1970s (Sandford et al. 1975) and was further developed at the Laboratory during the 1980s (Wohletz et al. 1984; Horn 1989) and 1990s, as summarized by Valentine (1998). Ongoing development by Italian volcanologists (Neri et al. 2003; Todesco et al 2002; Ongaro et al. 2002) and others applies multiphase theory to

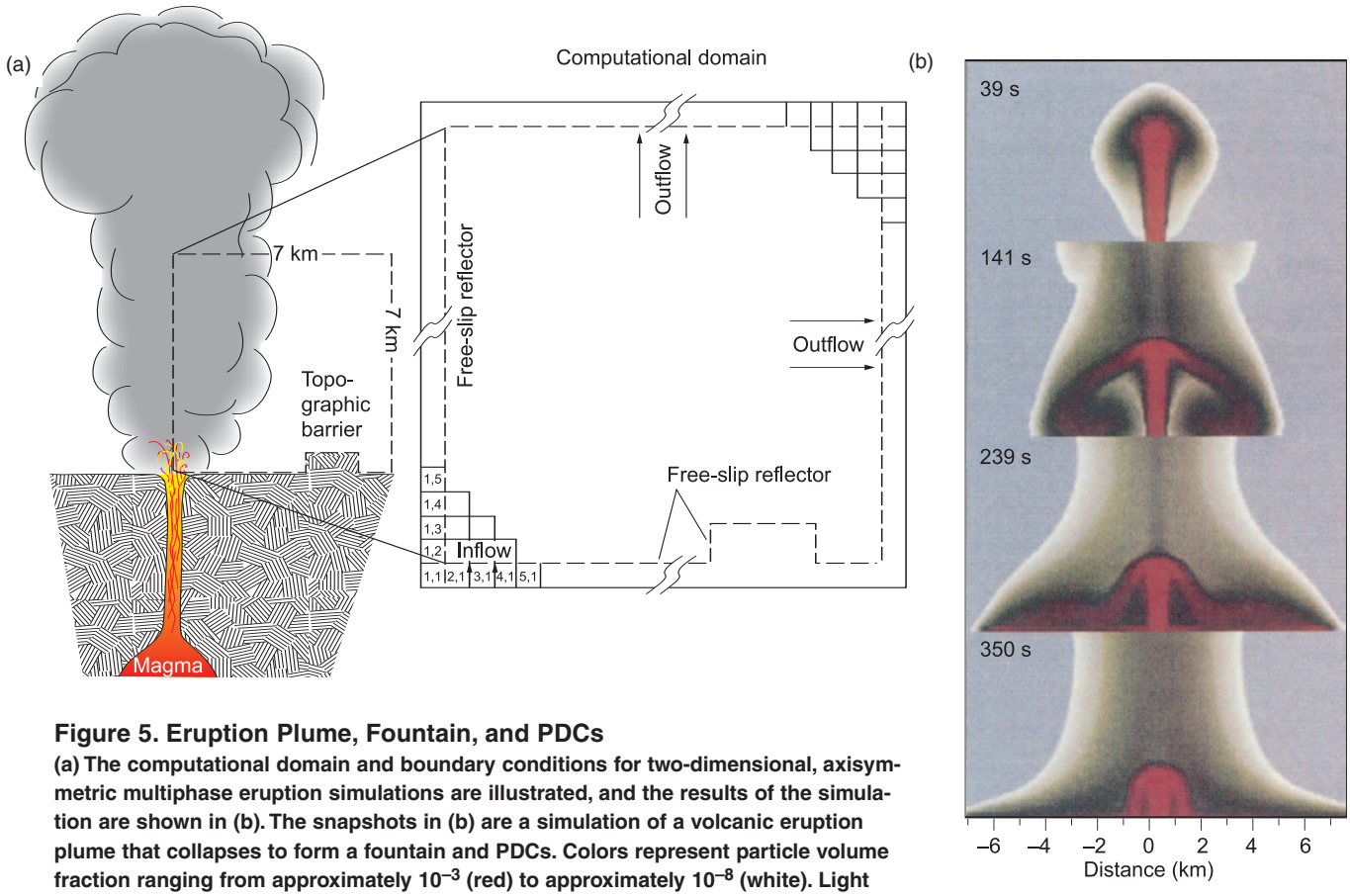


Figure 5. Eruption Plume, Fountain, and PDCs

(a) The computational domain and boundary conditions for two-dimensional, axisymmetric multiphase eruption simulations are illustrated, and the results of the simulation are shown in (b). The snapshots in (b) are a simulation of a volcanic eruption plume that collapses to form a fountain and PDCs. Colors represent particle volume fraction ranging from approximately 10^{-3} (red) to approximately 10^{-8} (white). Light blue represents "clean" ambient atmosphere. Vertical and horizontal scales are each 7 km. (Figure is adapted from Valentine et al. 1992 courtesy of the Geological Society of America.)

predict hazards to urban areas such as Naples, Italy, and to better understand the transport and deposition processes of PDCs. As in the conduit fluid model, the multiphase modeling of eruption plumes and PDCs computes the motion of a continuous, compressible gas phase (a mixture of erupted volatiles and entrained air) and one or more particle fields, as if they are interpenetrating fluids. In other words, the gas and particles are each treated as a fluid field, occupying the same volume according to their individual volume fractions (which must sum to unity). Each of these fields has an accompanying set of mass, momentum, and energy conservation equations. The fields can be coupled together by mass exchange, drag (momentum exchange), and heat exchange along with heat generated by drag. This multifield approach is valid only for problems in which the control volume (or representative elementary volume) is sufficiently large for particle behavior to be described as a field, rather than by each particle's dynamics. Valentine (1994) presented a multifield framework for a wide range of volcanic processes, including plumes and PDCs.

Figure 5 illustrates the results of a two-dimensional, time-dependent multiphase calculation (Valentine et al. 1992). This calculation, which would now be considered a first-generation multiphase volcano calculation, was axisymmetric (the symmetry axis is in the center of the snapshots—in reality, only a half-space calculation was done, and the results were "reflected" for the purposes of illustration). It accounts for one particle size and one gas species, and it has a regular, uniform grid (100 × 100 meters). A mixture of hot (1200 kelvins) gas and particles with an initial velocity of 290 m/s and gas pressure of 0.1 megapascal (equal to ambient) is injected into the atmosphere. The mass fraction of gas (water vapor) at the "vent" is 1.7%. Colors in the figure indicate particle volume fraction ranging from a high of about 10^{-3} (red) through black and white to a low of 10^{-9} (blue—relatively "clean" ambient atmosphere). The jet rises to an initial height of approximately 4 kilometers, at which point its initial momentum or kinetic energy is spent. Because the mixture is denser than the surrounding air at that point, the bulk of the material collapses to form a fountain while a dilute plume continues to rise above the eruption. At the spot where the collapsing mixture impacts the ground, it flows both outward and

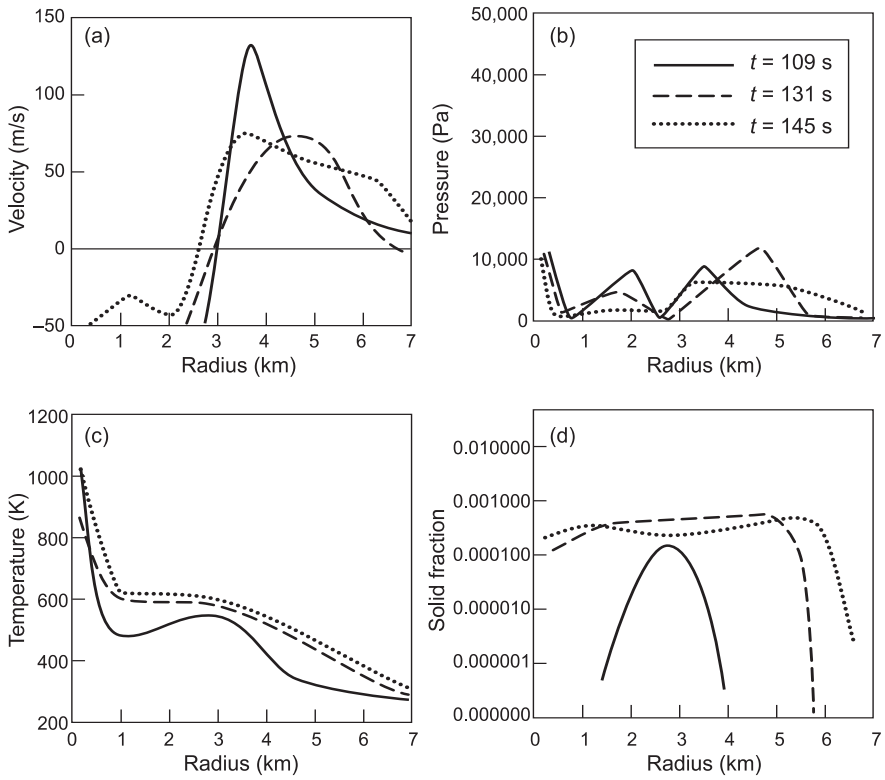


Figure 6. Damage-Causing Parameters Resulting from PDCs
Plots of radial (a) velocity, (b) dynamic pressure, (c) particle temperature, and (d) particle volume fraction are "measured" along the ground in a two-phase eruption simulation. These parameters are important for predicting hazards (for example, to buildings or people) in the area affected by PDCs. (Figure is adapted from Valentine and Wohletz 1989.)

ventward as a PDC. The ventward-flowing material is recycled into the eruptive jet, reducing the jet's vertical momentum and causing the fountain to decrease in altitude. The outward flowing material moves at velocities of several tens of meters per second, in a manner that varies with time as the overall dynamics evolve.

While the general fluid dynamics of these eruptions are of interest from a research perspective, in this section, we focus on parameters that relate to potential damage to structures on the ground. These parameters are flow temperature, velocity, and particle concentration. Flow velocity and particle concentration (through its effect on flow density) determine the dynamic pressure, P_{dyn} ($P_{dyn} = 1/2\rho u^2$, where ρ is the density of the mixture and u is the horizontal component of velocity), experienced by any object in the flow path. As an example, Figure 6 shows these parameters along the ground for three different times in a calculation similar to that discussed above. Figure 6(a) indicates that, as the PDC flows away from the point where the fountain impacts the ground (near the point where flow speeds cross from negative, or ventward-flowing, velocities to positive, or outward-flowing, velocities), it initially attains peak values approaching 150 m/s. As the flow field evolves, the peak PDC velocities decrease to about 70 m/s, and the radial distribution of velocity changes. Dynamic pressure—refer to Figure 6(b)—evolves through time as well, with values ranging from 5 to 10 kilopascals, spreading outward radially as the flow evolves. Temperatures on the ground—see Figure 6(c)—evolve toward a radially decreasing pattern, reflecting progressive heat transfer from particles and mixing with cooler atmosphere. The volume fraction of particles along the ground—shown in Figure 6(d)—stabilizes at about $1 - 2 \times 10^{-4}$ during this simulation. Results such as those illustrated in Figure 6 can be combined with information on the response of buildings to elevated temperature and dynamic pressure, for example, to predict damage from an eruption.

There have been a number of important advances in multifield modeling approaches for explosive eruptions over the past decade, most of which are described in Neri et al. (2003), Dartevelle (2004) and Dartevelle et al. (2004). Among them are the following: variable meshes that provide much better resolution for dynamics adjacent to boundaries such as the ground surface, where particle settling can produce steep gradients in flow properties and terrain can be represented; large-eddy simulation turbulence model; constitutive models that account for momentum transfer by particle collision whenever solid volume fraction is sufficiently high; capability for n particle classes (determined, for example, by size and/or material density), each represented by a set of mass, momentum, and energy field equations; and multiple gas species (for example, steam, air, or carbon dioxide). Using these new capabilities, Todesco et al. (2002) and Ongaro et al. (2002) are predicting values of damage-producing parameters for potential eruptions of the Vesuvius volcano in Italy that could endanger the heavily urbanized surroundings.

Structural Damage

The next step in predicting risk from explosive eruptions is to quantify the effects on people, buildings, and other infrastructure; here, we will focus on buildings. When exposed to a PDC, buildings can be damaged by thermal effects, high static pressures, dynamic pressure, projectiles (for example, large rocks or debris from upstream buildings), and potential burial by depositing particles. Thermal effects depend on the temperature conditions within the PDC, ignition conditions, and availability of oxygen for combustion. Data on the effects of projectiles on buildings are being compiled. Sources used are observations from recent eruptions as well as damage caused by debris from tornadoes and hurricanes.

Dynamic pressure from a passing PDC produces a lateral load on a building. Simple estimates of dynamic pressures produced by PDCs indicate that P_{dyn} could range from as high as approximately 10 megapascals (for a PDC with velocity of 300 m/s and particle volume concentration of 0.5) to approximately 1 kilopascal for a dilute, relatively slow current (velocities of a few tens of meters per second, particle volume concentrations of about 10^{-4}). Most buildings will experience severe damage with lateral loads of about 8 to 40 kilopascals—1 to 5 pounds per square inch—depending upon the type of construction (see, for example, Glasstone and Dolan 1977). Clearly, based on the reasonable range of P_{dyn} given above, many PDCs will totally destroy any buildings in their paths, and there is no point in understanding the details of structural response in regimes above approximately100 kilopascals, except for extremely strong monumental buildings. Nevertheless, many PDCs may result in lateral loads that would be expected to produce partial damage; however, even for the most

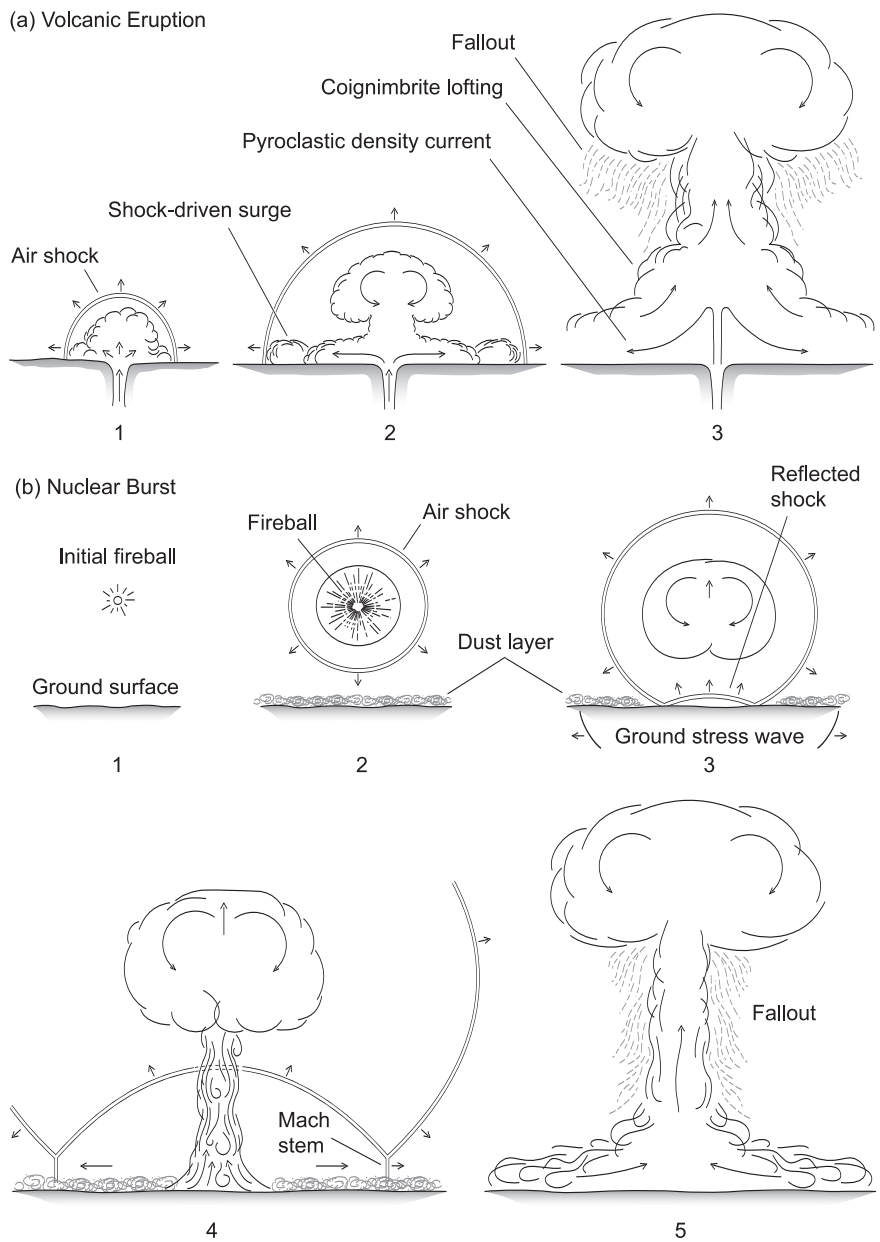


Figure 7. Comparing Volcanic Eruptions and Low-Altitude Nuclear Explosions
(a) An explosive volcanic eruption may generate an air shock because of the decompression of volcanic gases and the impulse of material flowing into the atmosphere. As the explosion grows, shock waves may drive a surge of particle-laden gas along the ground. Finally, as the eruption continues, the particle gas mixture may behave like a fountain, with PDCs flowing along the ground and a buoyant plume rising above the vent from which particles deposit by fallout. (b) A low-altitude nuclear explosion generates an air shock from the rapidly expanding fireball. An outward-moving Mach stem shock forms at the intersection of the incident and reflected air shock. As the fireball continues to expand, it also begins to rise; entrainment of ground debris into the rising fireball produces the characteristic fireball. Blast damage on the ground is caused by the Mach stem shock, which produces short-lived lateral forces on any structure in its path. (Adapted from *Journal of Volcanology and Geothermal Research*, 87, G. A. Valentine 1998, pp. 117–140 with permission from Elsevier.)

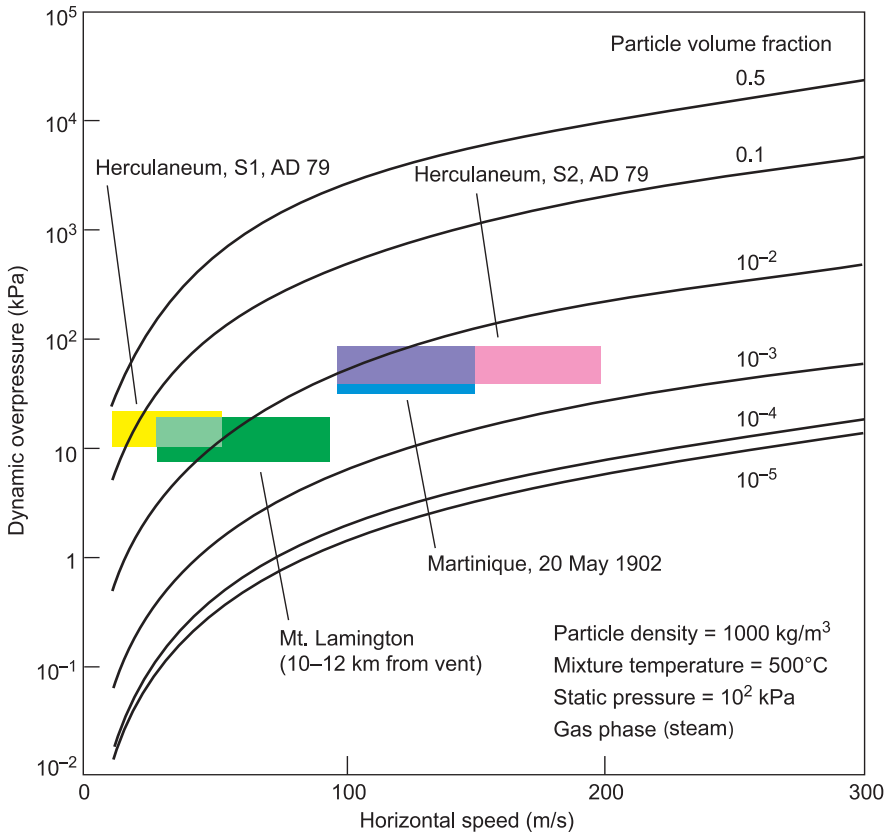


Figure 8. Dynamic Pressure as a Function of Velocity for Different Particle Volume Fractions

This plot shows dynamic pressure from PDCs as a function of velocity for different particle volume fractions. The colored boxes represent regimes for four historical eruptions. A combination of data was used, including structural damage and damage criteria from nuclear tests, as a proxy for PDC-induced damage. Although the quality of information on damage produced by historical eruptions is imprecise, the range of PDC speeds and particle concentrations that can be estimated from the damage includes consistent values that are suggested by, for example, evidence from sediment transport theory. (Adapted from *Journal of Volcanology and Geothermal Research*, 87, G. A. Valentine 1998, pp. 117–140, with permission from Elsevier.)

damaging PDCs, there will be zones around their margins, where conditions are not so severe. Understanding these factors is important for emergency mitigation and response planning in regions that are vulnerable to PDCs.

Interestingly, similar issues faced civil defense planners in the early years of the Cold War, but they were related to damage caused by nuclear weapons (eventually, with the adoption of the strategy of mutually assured destruction and large-yield fusion weapons, the details of damage to cities for civil protection became more or less moot). During those years, full-scale tests were conducted, whereby real buildings were exposed to nuclear blast loading; it is possible to use the structural response information from those tests as rough analogs for conditions in PDCs. Figure 7 illustrates the phenomena associated with an explosive eruption and a low-altitude nuclear burst. In a volcanic eruption, initial decompression of the erupting gas-particle mixture into the atmosphere can drive a shock wave that expands outward into the air. This might be followed by a blast-driven surge and, eventually, by full-scale PDCs that are of interest here. In a low-altitude nuclear burst, the expanding fireball pushes a strong air shock that expands spherically until it intersects the ground. The shock is then reflected upward from the ground, and a vertically oriented "Mach stem" shock forms at the intersection between the reflected and the incident shocks as the Mach stem continues to move outward. As it passes over a structure, the Mach stem creates a lateral load by two processes: (1) "diffraction" loading, which occurs as the shock is passing over the structure and the upstream side of the structure experiences a high pressure while the downstream side is still at ambient pressure; and (2) dynamic pressure loading, after the shock has passed and the building is subjected to a strong outward wind. All of this takes place in a very short time (seconds) in a nuclear case. In the volcanic case of PDCs, lateral loading is almost entirely due to dynamic pressure from the particle-laden flow, and that might be sustained for much longer times than in the nuclear case. In the absence of detailed data on damage from PDCs, however, it is reasonable to use nuclear effects data as a starting point.

Figure 8 shows the range of dynamic pressure as a function of PDC speed for several values of particle loading (volume fractions ranging from 10^{-5} to 0.5). Superimposed on these curves are boxes that represent the range of possible conditions as inferred from comparisons of nuclear effects data with observations of damage from four historical PDCs: the 1951 eruption of Mt. Lamington in Papua, New Guinea (Taylor 1958); the 1902 eruption of Mt. Pelee in Martinique (Lacroix 1904); and two PDCs that damaged the town of Herculaneum during the 79 AD eruption of Mt. Vesuvius. The height of each box represents our best estimate

of the possible range of dynamic pressures that could account for observed damage. The length of each box represents the range of PDC velocities as constrained by observations (or, in the case of Herculaneum, inferred from the characteristics of the deposits). Indirect information on the possible range of particle concentrations in these PDCs is consistent with the conditions indicated by the boxes.

The work described above served as a useful starting point for determining how structures respond to PDCs. In the past few years, there have been important new advances in observational data (mainly from the island of Montserrat, where PDCs that flowed out over residential areas were observed and the resulting damage was carefully documented—Baxter et al. (in press) and theoretical studies (Nunziante 2003). As a result, our understanding of PDC-induced damage is growing rapidly. This recent work indicates that the nuclear effects data, as applied by Valentine (1998), underestimates the damage caused by real PDCs for a given dynamic pressure. This greater damage results from several factors, such as shadowing or channeling effects by nearby structures, PDCs lasting longer than nuclear blasts, projectiles in the flows (particularly those derived from buildings just upstream), and heat. It is interesting to note that these results might, in turn, be used in studies of nuclear effects because there is now a great deal of interest in effects of low-yield devices in densely developed urban areas (for example, a terrorist device in a major city).

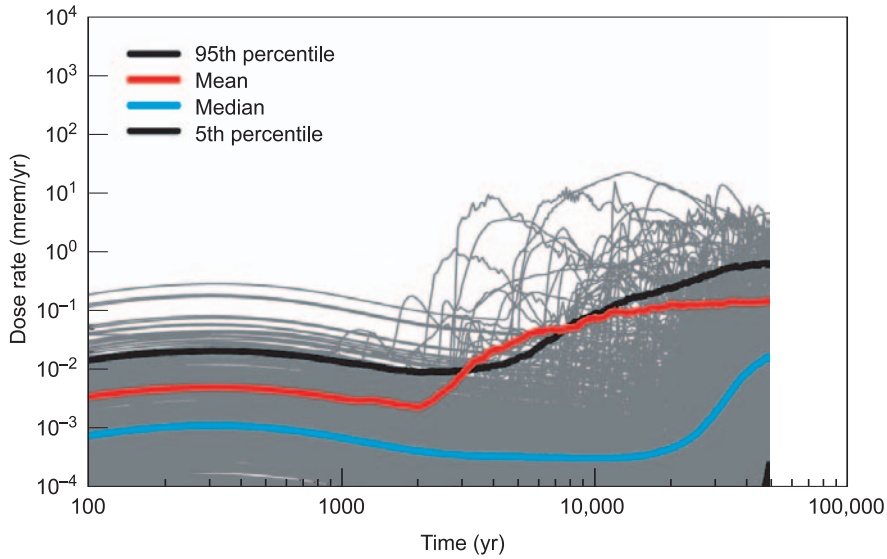


Figure 9. Predictions of Radiation Dose for Yucca Mountain
Predictions of radiation dose from a volcanic eruption are for a population located 18 km south of the proposed Yucca Mountain, Nevada, radioactive waste repository. The predictions are weighted by the probability of such an eruption, and they include processes such as waste entrainment into eruptive conduits, dispersion into the atmosphere and subsequent fallout, and contamination of ground water by damaged waste packages that remain in the underground environment. Gray curves indicate individual realizations of the integrated models. These predictions were for the Site Recommendation in 2001 (U.S. Department of Energy 2001) and are being updated for the December 2004 license application.

Examples of Applications

The predictive-earth-sciences framework plays an important role in addressing many problems of national importance: repository science necessary for closing the loop on current and future nuclear-fuel cycles; water resources research aimed at predicting the impacts of climate change and water usage on resource availability; sequestration of excess CO_2 into underground reservoirs to counter global warming due to use of fossil fuels; homeland security issues that involve interaction between terrorist events, the environment, humans, and infrastructure; and nuclear weapons effects from targeting, military vulnerability, and homeland vulnerability perspectives.

Predicting Volcanic Risk at Yucca Mountain. The research discussed above is guided by and fit together through the ultimate need to produce integrated predictions of the risk to humans who live around explosive volcanoes. One application of volcanology, in which the predictive-earth-sciences framework has played an especially strong role, is predicting the radioactive dose that might result if a small volcano were to erupt through the proposed Yucca Mountain repository. Figure 9 (U.S. Department of Energy 2001) shows the rolled-up results of those models that account for a probability distribution for occurrence of a volcanic event, subsurface interaction between rising magma and the repository, and subsequent eruption of nuclear waste onto the surface. The results of these simple models were cast in terms of probability distributions and then sampled by a Monte Carlo approach to produce a large number of runs, sampling all the modeled processes, and represented by the gray curves in Figure 9. This figure shows the predicted dose as a function of time into

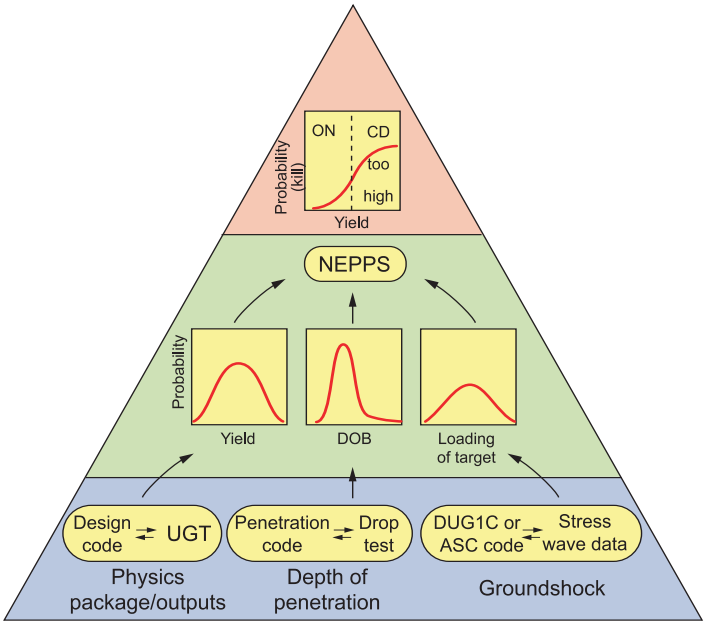


Figure 10. Predicting the Defeat of Deeply Buried Targets
This example of the predictive-earth-sciences framework is used for predicting the defeat of deeply buried targets by nuclear earth-penetrator weapons. Some of the components of the overall predictive system that need to be understood are the following: (1) the weapon outputs (neutrons, gamma rays, and x rays), which are predicted by a design code in conjunction with test data (either from historical underground nuclear tests or from other tests); (2) the depth of warhead penetration into the ground, which determines the amount of energy transmitted into the ground and is generally predicted on the basis of solid mechanics codes coupled with drop test and other data; and (3) propagation of that energy as a shock through heterogeneous rock types to produce shock loading at the underground facility. Predictions of ground shock propagation integrate continuum mechanics codes from a package such as an ASC code package with data from underground explosives tests. Other components that are not illustrated include the accuracy of the weapon regarding the intended detonation point, the response of the underground target itself to shock loading, and potential collateral effects such as air blast and fallout. All these can be integrated, accounting for uncertainty in each component, through a simulation tool such as NEPPS (developed in the Systems Engineering and Integration Group at Los Alamos), to produce, for example, a prediction of the probability of rendering the target ineffective as a function of weapon yield.

the future, the primary criterion for determining whether the repository will perform as specified by regulations. At early times (the first 1000 years), the mean value is dominated by dose produced by eruption of waste and direct fallout onto a hypothetical population. At later times, the mean value is dominated by contamination of ground water because of magma-induced damage to waste packages. The predictions represented by Figure 9 are being superseded by new calculations that incorporate more detailed models of magma–repository interactions that will form part of the basis for a license application in December 2004.

Defeating Underground Targets. The predictive-earth-science framework can also be used to study the effects of nuclear weapons as applied to defeating underground targets (see Figure 10). Several processes are involved in defeating a deeply buried target with a nuclear weapon: delivery of the weapon to the target, penetration into the ground if it is an earth penetrator, performance of the nuclear physics package, coupling and propagation of energy as groundshock to the underground target, and response of the target itself. Each of these processes requires a physics-based understanding in order to capture the inherent uncertainties. Probability distributions of each process are then rolled up in a Monte Carlo approach such as NEPPS (for Nuclear Earth Penetrator Planning System, developed in the Systems Engineering and Integration Group at Los Alamos), to produce a high-level prediction that might take the form of a probability of target defeat (or some other combination of parameters).

Concluding Remarks

Finally, a focus on predictive earth sciences provides a driver for several classes of underpinning basic research. These include upscaling, coupling across chemical and physical regimes (for example, coupling global climate predictions to regional scales for water resources studies), stochastic processes, extreme events (such as weapon effects or natural disasters), and the effects of having humans in the loop in environmental processes. In general, as with the Stockpile Stewardship Program, predictive earth sciences involve predicting the performance of coupled, nonlinear, multiscale processes that involve materials whose properties are heterogeneous and imperfectly characterized, where much of the data on the full-system performance are historical. ■

Further Reading

Baxter, P. J., R. Boyle, P. Cole, A. Neri, R. Spence, and G. Zuccaro. The Impacts of Pyroclastic Surges on Buildings at the Eruption of the Soufriere Hills Volcano, Montserrat. (Submitted to *J. Volcanol. Geotherm. Res.*)

BSC (Bechtel SAIC Company). 2004. Biosphere model report. MDL-MGR-MD-000001, Rev. 01. Las Vegas, Nevada: Bechtel SAIC Company.

Dartevelle, S. 2004. Numerical Modeling of Geophysical Granular Flows: 1. A Comprehensive Approach to Granular Rheologies and Geophysical Multiphase Flows. *Geochem. Geophys. Geosyst.* **5** (8).

Dartevelle, S., W. I. Rose, J. Stix, K. Kelfoun, and J. W. Vallance. 2004. Numerical Modeling of Geophysical Granular Flows: 2. Computer Simulations of Plinian Clouds and Pyroclastic Flows and Surges. *Geochem. Geophys. Geosyst.* **5** (8).

Eckhardt, R., D. L. Bish, G. Y. Bussod, J. T. Fabryka-Martin, S. S. Levy, P. W. Reimus et al. 2000. Yucca Mountain—Looking Ten Thousand Years into the Future. *Los Alamos Science* **26**: 464

Fisher, R. V. 1999. *Out of the Crater*: *Chronicles of a Volcanologist*. Princeton, New Jersey: Princeton University Press.

Gable, C. W. 2000. Mesh Generation for Yucca Mountain. In *Los Alamos Science* Number 26, Vol. 2, p. 472.

Glasstone, S., and P. J. Dolan. 1977. *The Effects of Nuclear Weapons*. US Department of Defense and the Energy Research and Development Administration.

Horn, M. 1989. "DANIEL: A Computer Code for High-Speed Dusty Gas Flows with Multiple Particle Sizes." Los Alamos National Laboratory report LA-11445-MS.

Lacroix, A., 1904. *La Montagne Pelèe et ses Éruptions*. Paris, France: Masson et Cie.

Macedonio, G., F. Dobran, and A. Neri. 1994. Erosion Processes in Volcanic Conduits and Application to the AD 79 Eruption of Vesuvius. *Earth Planet. Sci. Lett.* **121**: 137.

Neri, A., T. E. Ongaro, G. Macedonio, and D. Gidaspow. 2003. Multiparticle Simulation of Collapsing Volcanic Columns and Pyroclastic Flow. *J. Geophys. Res.* **108** (B4): 2202.Nunziante, L., M. Fraldi, L. Lirer, P. Petrosino, S. Scotellaro, and C. Cicirelli. 2003. Risk Assessment of the Impact of Pyroclastic Currents on the Towns Located Around Vesuvio: A Non-Linear Structural Inverse Analysis. *Bull. Volcanol.* **65** (8): 547.

Ongaro, T. E., A. Neri, M. Todesco, and G. Macedonio. 2002. Pyroclastic Flow Hazard Assessment at Vesuvius (Italy) by Using Numerical Modeling. II. Analysis of Flow Variables. *Bull. Volcanol.* **64** (3–4): 178.

Perry, F. V., B. M. Crowe, and G. A. Valentine. 2000. Analyzing Volcanic Hazards at Yucca Mountain. *Los Alamos Science* **26**: 492.

Sandford II, M. T., E. M. Jones, and T. R. McGetchin. 1975. Hydrodynamics of Caldera-Forming Eruptions. *Geol. Soc. Am. Abstr. Programs* **7** (7): 1257.

Sigurdsson, H., S. Carey, W. Cornell, and T. Pescatore. 1985. The Eruption of Vesuvius in AD 79. *Natl. Geogr. Res.* **1** (3): 332.

Taylor, G. A. M. 1958. "The 1951 Eruption of Mount Lamington, Papua." In *Aust. Bur. Miner. Resour. Geol. Geophys. Bull.*, Vol. 38. Australian Geological Survey Organization: Canberra, Australia.

Todesco, M., A. Neri, T. E. Ongaro, P. Papale, G. Macedonio, R. Santacroce, and A. Longo. 2002. Pyroclastic Flow Hazard Assessment at Vesuvius (Italy) by Using Numerical Modeling. I. Large-Scale Dynamics. *Bull. Volcanol.* **64** (3–4): 155.

US Department of Energy. 2001. "Yucca Mountain Science and Engineering Report—Technical Information Supporting Site Recommendation Consideration." US Department of Energy, Office of Civilian Radioactive Waste Management report DOE/RW-0539.

Valentine, G. A. 2003. Towards Integrated Natural Hazard Reduction in Urban Areas. In *Earth Science in the City*: *A Reader*. Edited by G. Heiken, R. Fakundiny, and J. F. Sutter. Washington, DC: American Geophysical Union.

———. 1998a. Damage to Structures by Pyroclastic Flows and Surges, Inferred from Nuclear Weapons Effects. *J. Volcanol. Geotherm. Res.* **87** (1–4): 117.

———. 1998b. Eruption Column Physics. In *From Magma to Tephra: Modelling Physical Processes of Explosive Volcanic Eruptions*. Edited by A. Freundt, and M. Rosi. Amsterdam: Elsevier.

———. 1994. Multifield Governing Equations for Magma Dynamics. *Geophys. Astrophys. Fluid Dyn.* **78**: 193.

Valentine, G. A., and K. R. Groves. 1996. Entrainment of Country Rock During Basaltic Eruptions of the Lucero Volcanic Field, New Mexico. *J. Geol.* **104** (1): 71.

Valentine, G. A., and K. H. Wohletz. 1989a. Environmental Hazards of Pyroclastic Flows Determined by Numerical Models. *Geology* **17** (7): 641.

———. 1989b. Numerical Models of Plinian Eruption Columns and Pyroclastic Flows. *J. Geophys. Res.* **94** (B2): 1867.

Valentine, G. A., K. H. Wohletz, and S. W. Kieffer. 1992. Effects of Topography on Facies and Compositional Zonation in Caldera-Related Ignimbrites. *Geol. Soc. Am. Bull.* **104** (2): 154.

Wohletz, K. H., T. R. McGetchin, M. T. Sandford II, and E. M. Jones. 1984. Hydrodynamic Aspects of Caldera-Forming Eruptions: Numerical Models. *J. Geophys. Res.* **89** (B10): 8269.

For further information, contact Greg Valentine (505) 665-0259 (gav@lanl.gov).

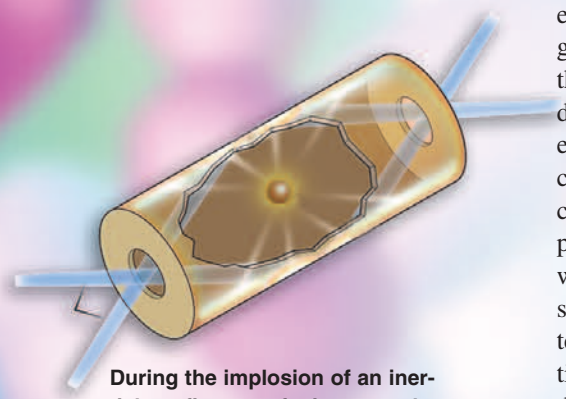
Quantum Molecular Dynamics

Simulating Warm, Dense Matter

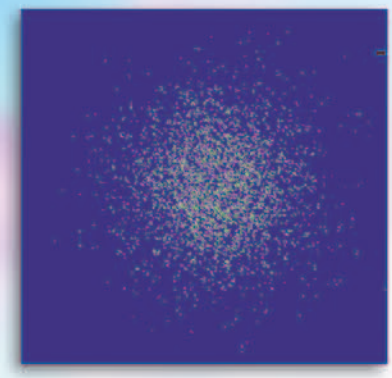
Lee A. Collins, Joel D. Kress, and Stephane F. Mazevet

Regions of warm, dense matter abound—from the interiors of giant gaseous planets, such as Saturn, and the atmospheres of white dwarf stars to laboratory plasmas in high-energy density generators and inertial confinement fusion capsules. Warm, dense matter, a sizzling "soup" of atoms, molecules, ions, and free electrons, is difficult to describe by standard techniques because it harbors multiple species and processes simultaneously—from ionization and recombination to molecular dissociation and association. Recently, quantum molecular dynamics (QMD), which can predict static, dynamical, and optical properties from a single, first principles framework, has been used to accurately predict properties of hydrogen, oxygen-nitrogen mixtures, and plutonium in the warm, dense state.

At the core of Saturn, there is warm, dense matter, ranging between 5000 and 6000 kelvins in temperature.



During the implosion of an inertial confinement fusion capsule, the deuterium-tritium atoms it contains become warm, dense matter.



This ultracold neutral plasma of electrons (green dots) and protons (pink dots) was created by heating microkelvin matter with a laser pulse. Surprisingly, this strongly coupled plasma has many features in common with warm, dense matter.

Warm, dense matter (WDM) appears in a wide variety of celestial and terrestrial environments—from the interiors of gaseous planets and atmospheres to the plasmas generated by high-energy-density machines and lasers. Other examples include shock-compressed cryogenically cooled materials, ultracold plasmas, and various stages in primary and secondary nuclear weapons. In general, these systems span temperatures from hundreds to tens of millions of kelvins and densities from about 1/100 to 100 times the density of a solid. The medium in each system resembles a "soup" of various species—including atoms, molecules, ions, and electrons—that exhibits distinctly nonclassical behavior in the interaction of all the particles. This material state is now generally referred to as warm, dense matter.

To model such systems, we have applied quantum molecular dynamics (QMD) simulation methods that treat the rapidly moving electrons quantum mechanically and the sluggish nuclei classically. In order to provide a diverse and systematic representation of the quantum mechanical effects, we have treated the electrons at various levels of sophistication—from a simple semi-empirical tight-binding model to a state-of-the-art finite-temperature density functional theory (DFT) approach. Because these techniques begin with only the most basic assumptions on the nature of the microscopic particle interactions from which all the macroscopic properties derive, they are designated as ab initio, or "from first principles." Through the QMD prescription, we can represent very complex structural and dynamical quantum processes that dominate these media. These methods currently allow for the treatment of a few hundred particles; however, scaling tests for massively parallel computers indicate that simulations with thousands of atoms will shortly become routine. An additional advantage of the QMD methods comes from their integrated nature. Because the QMD framework describes the elemental particle interactions, all the static, dynamical, and optical properties can be derived from an internally consistent set of principles, whereas in many models of dense media, the representations of these processes arise from different approaches at different levels of approximation.

The representation of a warm, dense system as an evolving sample of particles interacting through quantum mechanical forces has become possible only within the last two decades with the development of supercomputing capabilities. Los Alamos has pioneered this endeavor. As early as 1985, prototypical models, based on a semi-empirical determination of the quantum forces in representative snapshots of atomic configurations derived from classical molecular dynamics (MD) simulations, indicated the potential of such integrated approaches. By the mid 1990s, density functional methods had matured to the extent of effectively treating atomic samples of about 100 particles, a threshold for obtaining statistically significant macroscopic properties. This development initiated a renewed effort to employ these techniques together with semi-empirical approaches to model warm, dense systems. The ensuing years have witnessed the steady improvement of the basic quantum mechanical methods, efficient algorithms, and computational power, providing great accuracy in the characterization of these media.

Quantum Molecular Dynamics

A three-dimensional reference cell, containing N atoms (nuclei) at positions $\mathbf{R}[= \{\mathbf{R}_1 \ldots \mathbf{R}_N\}]$ and instanta-

neous momenta $\mathbf{P}[= \{\mathbf{P}_1 \ldots \mathbf{P}_N\}]$ and N_e electrons at $\mathbf{r}[= \{\mathbf{r}_1 \ldots \mathbf{r}_{Ne}\}]$, determines the basic working unit in QMD calculations. In order to represent the extended nature of the medium, we periodically replicate this cell throughout space and treat particle interactions both within the reference cell and with the repeated cells. The system evolves temporally according to a repeated two-step prescription. First, for fixed nuclear positions $\mathbf{R}(t)$ at a time t, we perform a sophisticated quantum mechanical calculation for the electrons. From the resulting electronic wave function $\Psi[\mathbf{r},\mathbf{R}]$, which depends only parametrically on the nuclear positions $\mathbf{R}$, we determine a force on each atom. Second, using this quantal force, we advance the nuclei over a short time δt by the classical equations of motion, yielding a new set of positions $\mathbf{R}(t + \delta t)$ and momenta $\mathbf{P}(t + \delta t)$ for the nuclei. This two-step process for temporally evolving a collection of particles forms the core of most MD approaches; the "Q" in QMD arises from the use of quantum mechanics to describe the electronic component.

We calculate the electronic many-body wave function $\Psi[\mathbf{r},\mathbf{R}]$ by solving the Schrödinger equation:

$$H\,\Psi[\mathbf{r},\mathbf{R}] = E\,\Psi[\mathbf{r},\mathbf{R}] \quad ,$$

where the Hamiltonian operator H has the form

$$H = T_e + V_{ee} + V_{eN} + V_{NN} \ .$$

T_e represents the kinetic energy of the electrons (e), and V_{ab} gives the interaction between the electrons (ee), the electrons and the nuclei (eN), and the nuclei (NN).

Numerous methods exist for solving the Schrödinger equation; however, the two most popular either construct the multicoordinate state function Ψ directly or employ the electron density $n(\mathbf{r}_i)$, determined by integrating the probability density $|\Psi|^2$ over all but one spatial variable. This density depends on only a single point in space and forms the basis of density functional theory (DFT). Both approaches usually invoke decomposition into single-electron orbitals, which in turn are expanded in a basis of simple functions such as Gaussians or plane waves. This reduction transforms the Schrödinger equation into a matrix eigenvalue problem for which powerful iterative techniques can effectively produce solutions.

We employ the wave function approach to generate an efficient, approximate procedure for solving the Schrödinger equation and rapidly advancing the molecule dynamics equations. In this tight-binding technique, we replace the matrix elements of H with semi-empirical forms fit to experimental data and specific theoretical results. The method still contains all the important processes afoot in the WDM regime: molecular dissociation and association, ionization and recombination, as well as electron collision and attachment.

In DFT, we use a very accurate form of these matrix elements that includes all the basic electrostatic and quantum effects (exchange and correlation). Although more expensive to calculate, this formulation yields a very accurate representation of a many-electron system. We typically use an extended form of DFT that includes finite temperature effects and some nonlocality (generalized gradient approximation). Most simulations evolve in local thermodynamical equilibrium with the electronic and nuclear kinetic temperatures set equal.

As indicated, for each MD time step, we obtain a set of positions and momenta $\{\mathbf{R}\,(t),\mathbf{P}(t)\}$ for the nuclei and the quantum mechanical state function $\Psi[\mathbf{r},\mathbf{R}]$ for the electrons. From this information, we can determine static properties, such as pressure and internal energies, and thus the equation of state (EOS), as well as dynamical properties such as diffusion, viscosity, order parameters, and thermal conductivities. On the other hand, the state function for the electrons yields optical properties such as electrical conductivity, reflectivity, dielectric functions, and opacities. The derivation of these properties from an internally consistent set of basic quantities $[\mathbf{R},\mathbf{P},\Psi]$ highlights a critical advantage of the QMD formulation.

The large variety of systems and environments explored by the QMD approaches attests to their great flexibility and applicability. Such applications include the following: isotopic plasma mixtures of dense hydrogen, nitrogen, and oxygen; highly compressed rare-gas solids, alkali metals near melt and along the vapor-liquid coexistence boundary, impurity atoms in dense hydrogen plasmas, shock-compressed liquids of atoms and hydrocarbons, and disorder in semiconductors.

For some of these cases, detailed experimental data exist. The generally good agreement obtained with the results of QMD simulations provides an effective validation of the technique across an extensive range of conditions and media. This validation proves particularly critical for the deployment of QMD into regimes of matter under extreme conditions, totally inaccessible to current experiments but vitally important to many national missions.

As a demonstration of the efficacy of these methods, we consider several representative examples.

Static Properties: Equation of State

The EOS of a material gives pressure and internal energy as a function of density and temperature and forms a basic component of any

model of a macroscopic system. The shock Hugoniot experiment provides the principal means of probing equations of state. In this case, a well-characterized shock is driven through a medium by the impact of a flier plate (from either a gas gun or a high-energy density device) or of a high-intensity laser pulse. The shock pressures (P), densities (ρ), and specific internal energies are determined by the Rankine-Hugoniot jump conditions across the thin shock front. These equations relate flow velocities and thermodynamical variables in the shocked state to those in the initial state. Therefore, knowing the initial conditions and the EOS, we can determine the value of the pressure—for example, at the final conditions—and compare with experimental observations.

As a first example, we focus on hydrogen, both for its deceptive simplicity and for the considerable controversy that has raged over its EOS. From an atomic physics standpoint, hydrogen, having but one proton and one electron, is the simplest element known. Surprisingly, its phase diagram displays considerable complexity. For a temperature range between 10^4 and 10^5 kelvins and a density range between 0.1 and 1.0 gram per cubic centimeter (g/cm^3), hydrogen exists as a dense diatomic fluid. As the temperature and density increase, the fluid undergoes continuous dissociation and ionization to become a fully ionized plasma consisting solely of free electrons and protons. For this regime, the challenge in obtaining a meaningful EOS lies in accurately describing the evolution of the delicate balance of atomic, molecular, and ionized species constituting the fluid. This complicated nature of hydrogen becomes evident in Figure 1, which displays the electronic probability density around the nuclei for a snapshot within a QMD simulation.

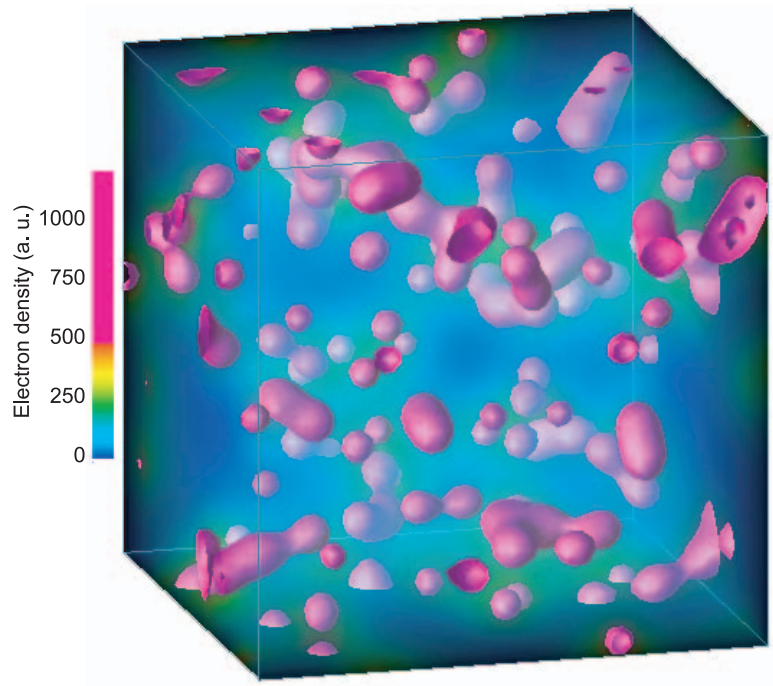


Figure 1. QMD Simulation of Hydrogen Electronic Density
A three-dimensional snapshot of a representative cubic sample of 128 highly compressed hydrogen atoms (1 g/cm^3 at 29,000 K) at a particular time within the QMD simulation shows the probability of finding an electron at a particular location. Magenta indicates the highest probability; blue, the lowest. The electron density indicates the presence of molecular systems (magenta ellipsoidal structures), atomic systems (magenta spheres), and free electrons in the intervening space.

The EOS of hydrogen and its isotope deuterium has received new attention because of recent laser experiments that seemed to call into question older models. Figure 2 displays the current status and depicts the pressure as a function of density for the shock compression (or the Hugoniot) of a molecular deuterium sample that was, at first, cryogenically cooled. Until the laser experiments, the established Hugoniot came from a chemical model in the SESAME equation of state tables compiled at Los Alamos in the 1970s, which yielded a maximum compression $\eta = 4$, given by the ratio of the density ρ to the initial density of the sample ρ_0. Experiments at the NOVA laser facility of the Lawrence Livermore National Laboratory (Lawrence Livermore) indicated a far more compressible medium with $\eta = 6$. This difference has profound ramifications for such diverse fields as planetary interiors (refer to Saumon and Guilot 2004) and nuclear weapons.

To gain insight into this experimental disagreement, we performed QMD calculations using the simple semi-empirical, tight-binding MD method and the very sophisticated density function approach (DFT-MD). Our QMD values agreed much better with the SESAME results and with subsequent, similar ab initio calculations—for example, the Path Integral Monte Carlo (PIMC). This agreement between the QMD and PIMC results has an

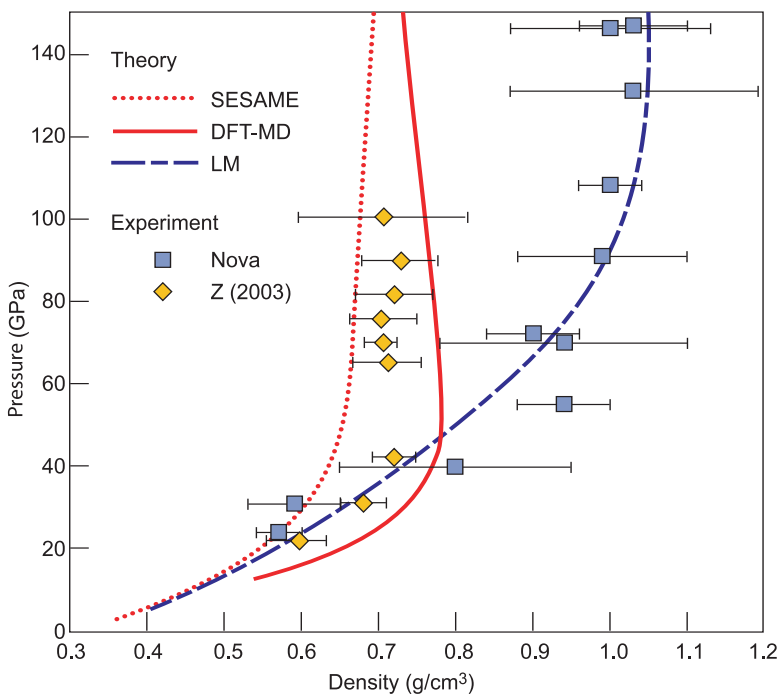


Figure 2. The Principal Hugoniot of Deuterium
The curves in this plot represent theoretical results from QMD simulations (DFT-MD) and from two free-energy minimization models, one used in the SESAME tables and the other based on linear mixing (LM). The experimental results are from the NOVA laser at Lawrence Livermore and the Z-machine at Sandia in New Mexico.

additional poignancy in that the PIMC treats the electronic and nuclear interactions by an approach completely independent of DFT. In 2000, a new set of experiments at the Sandia National Laboratories (Sandia), employing a flier plate accelerated by a pulse-power machine (Z-machine), produced results in close accord with the ab initio methods. Finally, within the last two years, Russian experiments (conducted at Sarov) with explosively generated converging shock waves supported the Sandia findings. The final verdict on the EOS of hydrogen awaits further experimental trials; however, the good agreement between the ab initio MD simulations and the experiments at the Z-machine and in Russia gives a strong penchant for the stiffer compressibility of hydrogen.

As a second example of the broad applicability of these approaches, we present comparisons of QMD simulations for compressed molecular nitrogen, oxygen, and nitrogen oxide (NO), which are similar to hydrogen in some respects: They all have molecular liquid states at very low temperatures, large dissociation energies, and moderate ionization potentials. For these three species though, several gas-gun experiments have probed a larger span of the Hugoniot than for hydrogen. Figure 3 displays the excellent agreement obtained between the QMD simulations and the experiments for the three species along the principal Hugoniot. This agreement indicates that the QMD approach can accurately characterize the progress of a very complex medium through many different stages. For example, the kink seen to varying degrees in all three Hugoniots indicates the transition from a molecular to an atomic fluid.

Experiments rarely follow the full evolution of a system but usually provide only final conditions based on prescribed starting values. On the other hand, QMD simulations can continuously monitor the state of the medium, yielding such valuable information as its constitution. To obtain better insight into the temporal development of a stressed medium, we examined, as a representative system, NO during compression from a cryogenic molecular liquid to a warm, dense fluid. The pair correlation function $g_{\alpha\beta}(r)$, which gives the probability of finding particles of type β a given distance r from a reference particle of type α, serves as an effective tool for tracing the change in the constituents. Figure 4(a) shows the conditions near the start of the compression in which the system consists mainly of NO molecules, as evidenced by the large peak in $g_{\mathrm{NO}}(r)$ at the average internuclear separation $R_{eq}(\mathrm{NO})$ for the molecular species. As both the density and temperature increase, the NO dissociates, and the freed nitrogen atoms combine into nitrogen molecules, but the oxygen remains in an atomic state. Figure 4(b) clearly depicts this behavior by the large peak in $g_{\mathrm{NN}}(r)$ near $R_{eq}(\mathrm{N}_2)$ and the very weak peak around $R_{eq}(\mathrm{O}_2)$. This finding challenges the present assumptions on modeling NO in overdriven shocks and has significant ramifications for explosives and high-pressure reactive chemistry.

Optical Properties

From the wave function that characterizes the electrons, we can also generate optical properties for the medium, including absorption coefficients, dielectric functions, reflectiv-

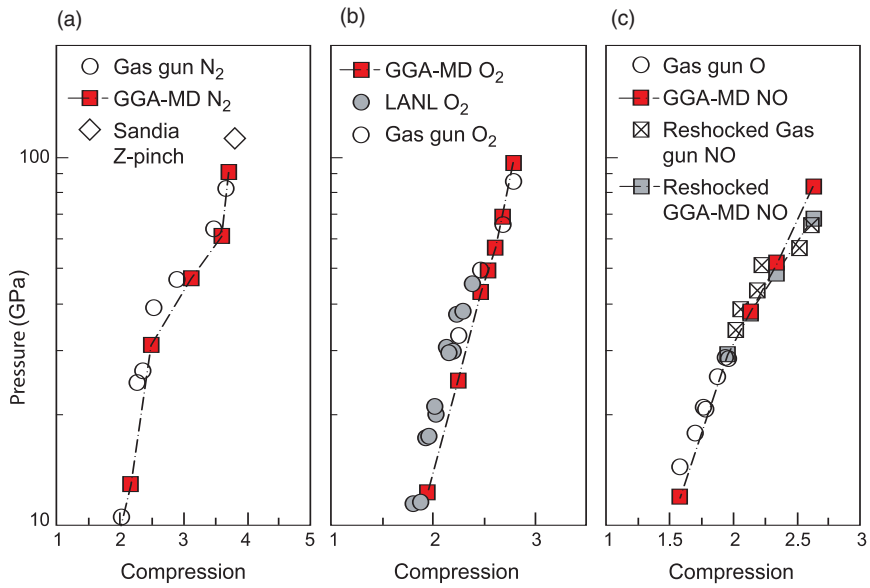


Figure 3. QMD vs Experiment for the Principal Hugoniot of N_2, O_2, and NO

Pressure is shown as a function of compression ($\eta = \rho/\rho_0$) along the principal Hugoniot for nitrogen (a), oxygen (b), and NO (c). Each panel compares QMD theoretical results (GGA-MD, red squares) with gas gun and Z-pinch experiments. In (c) reshock results are also shown. The excellent agreement between theory and experiment is noteworthy.

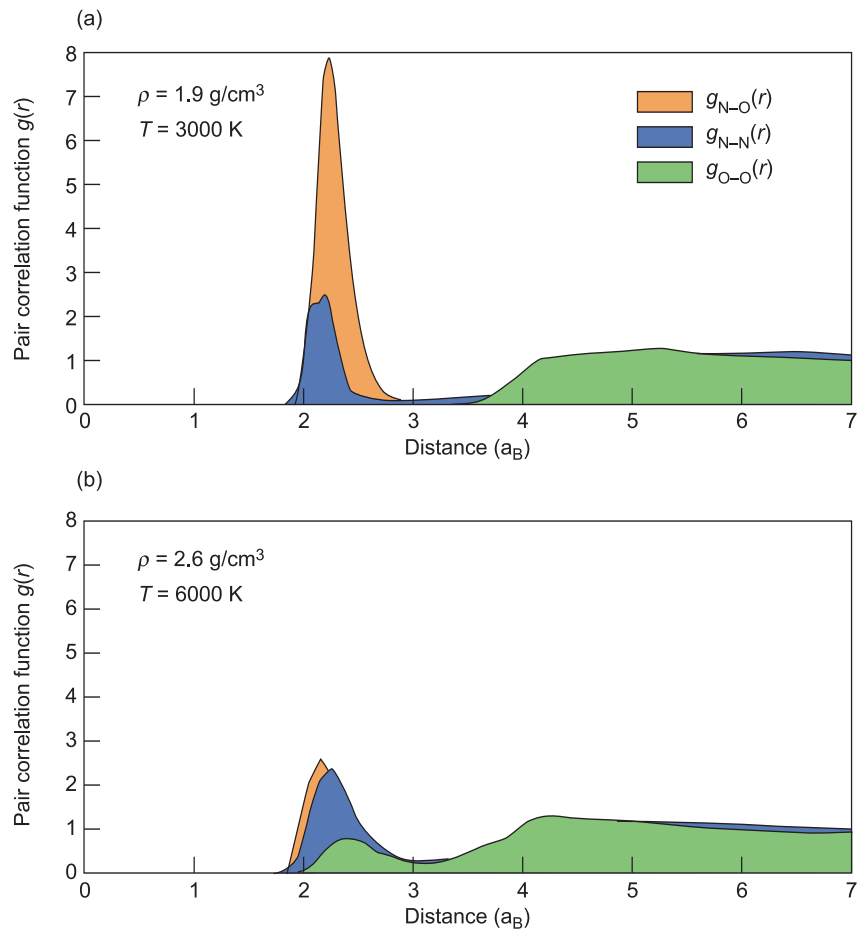


Figure 4. Pair Correlation Function for a NO Fluid under Shock Compression

The $g(r)$ pair correlation function gives the probability of finding an atom of a particular type a distance r from a reference atom. It therefore yields information about the composition of the fluid. The panels depict two sets of conditions: (a) density = 1.9 g/cm^3 and temperature = 3000 K, and (b) density = 2.6 g/cm^3 and temperature = 6000 K. The fluid begins as a pure system of NO molecules. As the temperature and density increase under compression, the NO dissociates and nitrogen molecules form. Oxygen remains almost entirely atomic.

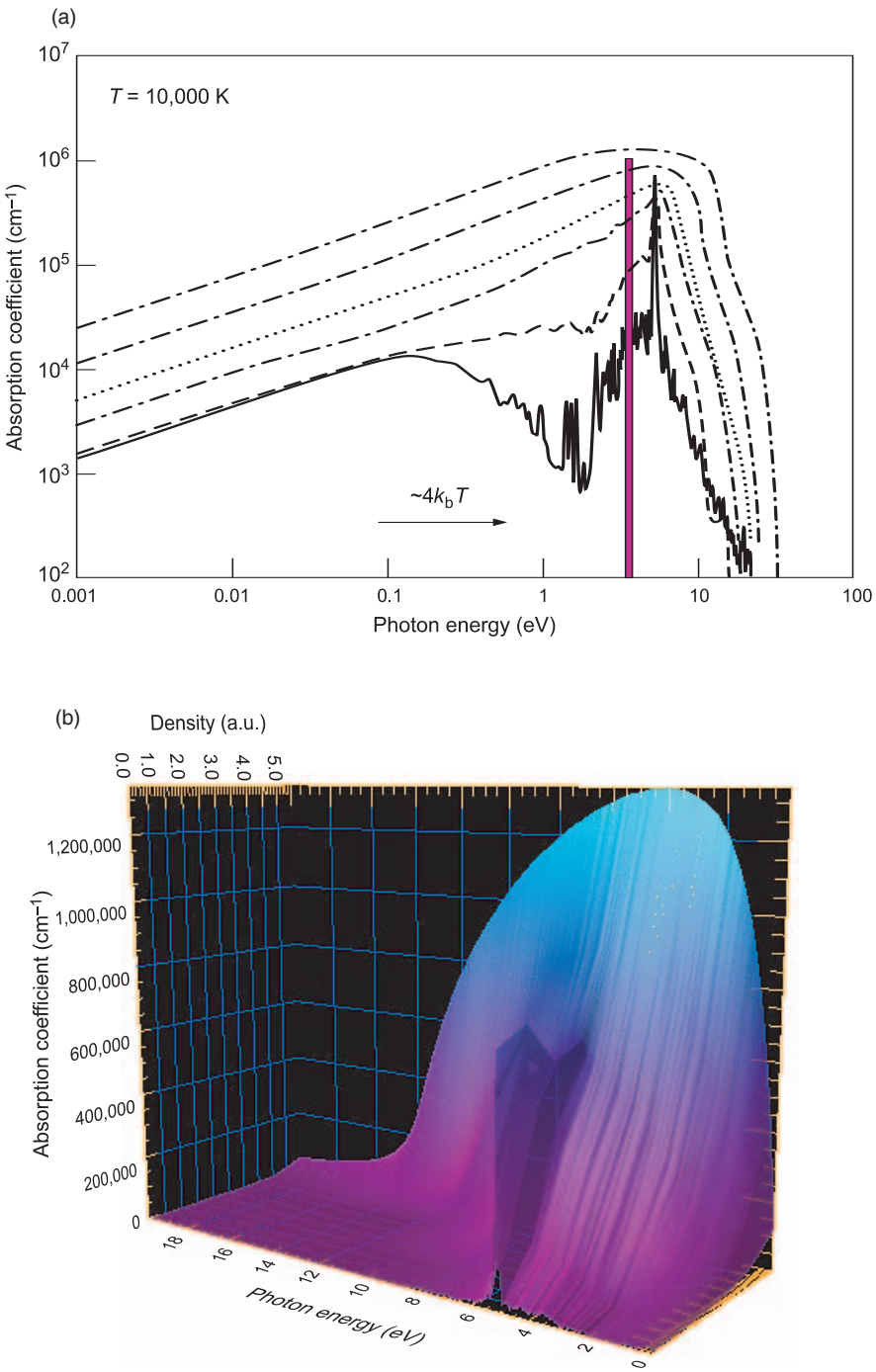


Figure 5. QMD Results for the Aluminum Absorption Coefficient
(a) From highest to lowest, the curves in this plot represent densities from 2.0 g/cm^3 to 0.025 g/cm^3 and show the trend in optical properties as the system moves from a solid to a gas. The absorption coefficient at the lowest density exhibits a distinct spectral line around 5 eV, originating from the 3s to 3p atomic transition. As the density increases, this feature broadens and melds with the continuous background. At the highest density, the profile resembles that of a dense metallic fluid. (b) The lower panel presents the same information but with a separate density axis.

ities, indices of refraction, and opacity. Opacity, a measure of the absorption of radiation in matter, is an important quantity in modeling diverse phenomena in astrophysics and in designing weapons. Many of the opacity libraries commonly used for standard macroscopic modeling programs (for example, in hydrodynamics) employ physical models that have not seen significant revision in decades. During this time, developments in a wide variety of fields, including weapons, inertial confinement fusion, high-energy density, and astrophysics, have required extensions of these libraries into new and complex regimes. Such an extension requires careful validation, either from experiments or from more-sophisticated theoretical methods, of the physical models that produce the opacity data. Since experiments have proved difficult within these new realms, as witnessed by the controversy over the EOS of compressed deuterium, ab initio simulation techniques, such as QMD, provide the best venue for making meaningful critiques of these models.

As indicated, these ab initio approaches produce a consistent set of material and optical properties from the same simulation. In contrast, the opacity libraries consist of a collection of approximate models. Therefore, an understanding of the differences in the opacities between the libraries and ab initio approaches requires a detailed examination of the underlying material properties, such as the EOS, and optical properties (absorption coefficient). To this end, we have performed large-scale QMD simulations of hydrogen and of aluminum and compared representative properties with the results from standard opacity libraries, in particular, the Light Element Detailed Configuration Opacity (LEDCOP) Code from Los Alamos.

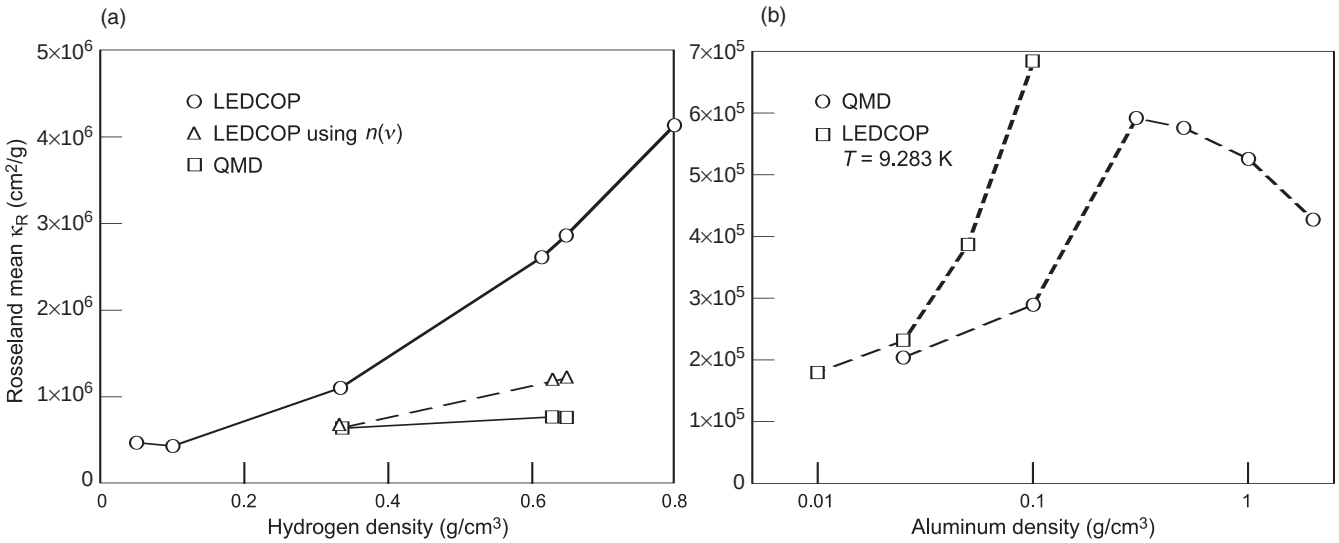


Figure 6. QMD- and LEDCOP-Derived Rosseland Mean Opacities for Hydrogen and Aluminum
We used QMD and LEDCOP to obtain Rosseland mean opacities as a function of density for hydrogen (a) and aluminum (b) at fixed temperatures of 48,000 K and 10,000 K, respectively. LEDCOP is based on an isolated atom perturbed by the surrounding medium; QMD considers all the atoms in the reference cell on an equal footing. "LEDCOP using *n*(*ν*)" is based on the isolated atom absorption coefficient and the QMD index of refraction. At low densities in the gas phase, LEDCOP and QMD agree well. As the density increases, the effects of the medium become pronounced, and the perturbative treatment fails.

The absorption coefficient $\alpha(\nu)$, which gives the attenuation of radiation as a function of frequency ν (photon energy) at a given density ρ and temperature T, is the fundamental physical quantity determined by both the QMD and LEDCOP. This quantity has a direct relationship to the frequency-dependent electrical conductivity $\sigma(\nu)$ and index of refraction $n(\nu)$ of the medium [$\alpha(\nu) = 4\pi\sigma(\nu)/n(\nu)$]. For zero frequency, $\sigma(\nu)$ yields the more familiar direct-current (dc) electrical conductivity σ_{dc}[$= \sigma(0)$], which determines the degree of current flow in a substance. Materials with σ_{dc} above 10,000 per ohm centimeter (Ω cm) are considered good conductors or metals; those with σ_{dc} below 1000 $(\Omega\ \text{cm})^{-1}$, quasi-metals, or near insulators. Finally, the inverse of $\alpha(\nu)$, integrated in frequency over the derivative of the normalized Planck function, yields the ubiquitous Rosseland mean opacity κ_R.

Figures 5(a) and 5(b) portray, from different perspectives, the absorption coefficient as a function of photon energy at a given temperature (10,000 kelvins) for aluminum, as the metal passes between two very different physical states: from a warm fluid at solid density (2.0 g/cm^3) to a gas (0.025 g/cm^3). The finite value of $\alpha(\nu)$ at low frequencies implies a conducting medium of varying degrees. At high densities, the associated σ_{dc} of about 30,000 $(\Omega\ \text{cm})^{-1}$ indicates a metallic fluid, whereas the small values—<1000 $(\Omega\ \text{cm})^{-1}$—in the gas phase signify an almost insulating medium. For the gas, we also note the appearance of structure in the absorption coefficient. The peak around 5 electron volts corresponds to the atomic line of neutral aluminum for the transition from 3s to 3p. This transition again identifies the low-density state as a collection of neutral atoms with a few free electrons. The bar at an energy corresponding to $4k_BT$ represents the regime with maximum contribution to the mean opacity.

In Figure 6, we compare QMD and LEDCOP calculations of the Rosseland mean opacity κ_R at a fixed temperature as a function of density for both hydrogen and aluminum. As the density decreases, the two approaches show better agreement. This agreement follows from the nature of the LEDCOP model, in which density effects enter only perturbatively upon an isolated atom. For higher densities, the strong overlap among the wave functions on different atomic centers obviates this perturbative view and needs a more democratic treatment of all the system electrons. This comparison demonstrates the ability of QMD simulations to validate and improve current opacity libraries as they are extended into new regimes.

Dynamical Properties

Plutonium may very well be the most complex of elements. At atmospheric pressure, the phase diagram shows six equilibrium solid phases as well as a liquid phase. A study of plutonium introduces another level of complexity because

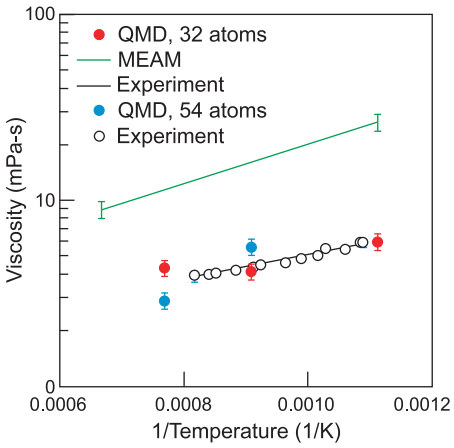


Figure 7. Temperature Dependence of Plutonium Viscosity

Molecular dynamics simulations are compared with a quantum mechanical (QMD) approach (red and blue squares) and a classical-potential (MEAM) approach (green line). QMD results are for samples of 32 atoms (red squares) and 54 atoms (blue squares). Experimental points are indicated by open squares. Quantum-derived forces appear important in determining the dynamic properties of this heavy system.

electron spin (magnetic behavior) must also be considered. Although the spin-DFT calculations for the face-centered-cubic (fcc) lattice structure (δ-plutonium) predict an antiferromagnetic (AF) state (in disagreement with the observations of a nonmagnetic state), the predicted structure is quite good, with an atomic volume (V) about 9 percent less than experiment. (An AF state has a net zero magnetic moment, with spin directions alternating up and down at each atomic site on the lattice.) We proceeded to study liquid plutonium with QMD in order to explore whether the quantum-derived forces provide a better description than does the interatomic potential of the classical Modified Embedded Atom Method (MEAM). We worked under the hypothesis that the predicted spin-DFT structural behavior will dominate over the predicted magnetic behavior, especially because the latter should be diminished by the disorder introduced in the liquid structure. An AF-like solution was found for the spin-DFT calculation (net zero magnetic moment with spins allowed to fluctuate on each atom during the MD trajectory). Radial distribution functions were calculated and self-diffusion coefficients (D) were derived from the mean-squared displacement of the atoms determined from the MD trajectory. In Figure 7, we compare QMD and classical MD (employing the MEAM potential) calculations of the viscosity of liquid plutonium with experimental data. In the classical MD calculations (performed by Los Alamos scientists Frank Cherne of the Materials Dynamics Group, Michael Baskes of the Structure/Property Relations Group, and Brad Holian of the Theoretical Chemistry and Molecular Physics Group), a 1024-atom simulation cell and nonequilibrium driven-slab boundary conditions were employed to compute the viscosity directly. In the QMD simulations, we calculated D from the mean-squared displacement of the atoms determined from an equilibrium MD trajectory and a 54-atom cell. The viscosity (η) was then calculated from a Stokes-Einstein relationship, namely, $(D\eta b)/(k_b T) = c$, where $b = V^{1/3}$ and k_b = Boltzmann constant. The dimensionless constant $c = 0.18$ ($\pm$ 0.02) is based on an analysis of experimental data for 21 different liquid metals (this work was conducted by Eric Chisolm and Duane Wallace of the Mechanics of Materials and Equation-of-State Group at Los Alamos). (A value of $c = 0.154 \pm 0.0123$ was determined from the classical MD simulations.) The MEAM potential was developed to describe the solid phases of plutonium; therefore, the results from the liquid simulations are a prediction. In this light, the MEAM values in Figure 7 are in fair agreement with experiment (within a factor of 5). The preliminary QMD results agree reasonably well with experiment, although we note that a 54-atom simulation is probably too small to provide a definitive answer.

Preliminary simulations at 1300 kelvins for (a) 54 atoms with the net zero magnetic moment relaxed and for (b) 108 atoms (requiring at least 32 Pentium processors in parallel) are yielding viscosities consistent with those illustrated in Figure 7. The relaxed magnetic moment result illustrates that the magnetic state apparently has little influence on the structure and dynamics of the liquid. With more computational horsepower, our goal is to calculate viscosity directly with QMD using nonequilibrium MD boundary conditions and 1024 atoms.

In summary, QMD simulations have proved an effective, versatile theoretical and computational approach to treating a large variety of warm, dense systems of particular interest to a broad number of Laboratory programs by providing systematic, integrated techniques for probing matter under extreme conditions. ■

Further Reading

Bickham, S. R., J. D. Kress, L. A. Collins, and R. Stumpf. 1999. *Ab Initio* Molecular Dynamics Studies of off-Center Displacements in CuCl. *Phys. Rev. Lett.* **83** (3): 568.

Collins, L. A., and A. L. Merts. 1985. Electronic Structure of Clusters of Atoms in a Dense Plasma. In *Proceedings of the 2nd International Conference on Radiative Properties of Hot Dense Matter*, *Sarasota, Florida*, *Oct. 31–Nov. 4*, *1983*. Edited by J. Davis, C. Hooper, R. Lee, and A. Merts, 385. Singapore: World Scientific.

Desjarlais, M. P., J. D. Kress, and L. A. Collins. 2002. Electrical Conductivity for Warm, Dense Aluminum Plasmas and Liquids. *Phys. Rev. E* **66** (2): 025401 (R).

Kress, J. D., I. Kwon, and L. A. Collins. 1995. Simulation of Impurity Line Shapes in a Hot, Dense Plasma. *J. Quant. Spectros. Radiat. Transf.* **54** (1–2): 237.

Kress, J. D., S. R. Bickham, L. A. Collins, and B. L. Holian. 1999. Tight-Binding Molecular Dynamics of Shock Waves in Methane. *Phys. Rev. Lett.* **83** (19): 3896.

Kwon, I., L. A. Collins, J. D. Kress, N. Troullier, and D. L. Lynch. 1994. Molecular Dynamics Simulations of Hot, Dense Hydrogen. *Phys. Rev. E* **49**: R4771.

Lenosky, T. J., J. D. Kress, L. A. Collins. 1997. Molecular-Dynamics Modeling of the Hugoniot of Shock Liquid Deuterium. *Phys. Rev. B* **56** (9): 5164.

Mazevet, S., L. A. Collins, and J. D. Kress. 2002. Evolution of Ultracold Neutral Plasmas. *Phys. Rev. Lett.* **88** (5): 055001.

Mazevet, S., P. Blottiau, J. D. Kress, and L. A. Collins. 2004. Quantum Molecular Dynamics Simulations of Shocked Nitrogen Oxide. *Phys. Rev. B* **69**: 224207.

Mazevet, S., L. A. Collins, N. H. Magee, J. D. Kress, and J. J. Keady. 2003. Quantum Molecular Dynamics Calculations of Radiative Opacities. *Astron. Astrophys. Lett.* **405**: L5.

Mazevet, S., J. D. Johnson, J. D. Kress, L. A. Collins, and P. Blottiau. 2002. Density Functional Calculation of Multiple-Shock Hugoniots of Liquid Nitrogen. *Phys. Rev. B* **65**: 014204.

Militzer, B., D. M. Ceperley, J. D. Kress, J. D. Johnson, L. A. Collins, and S. Mazevet. 2001. Calculation of a Deuterium Double Shock Hugoniot from *Ab Initio* Simulations. *Phys. Rev. Lett.* **87** (27): 275502.

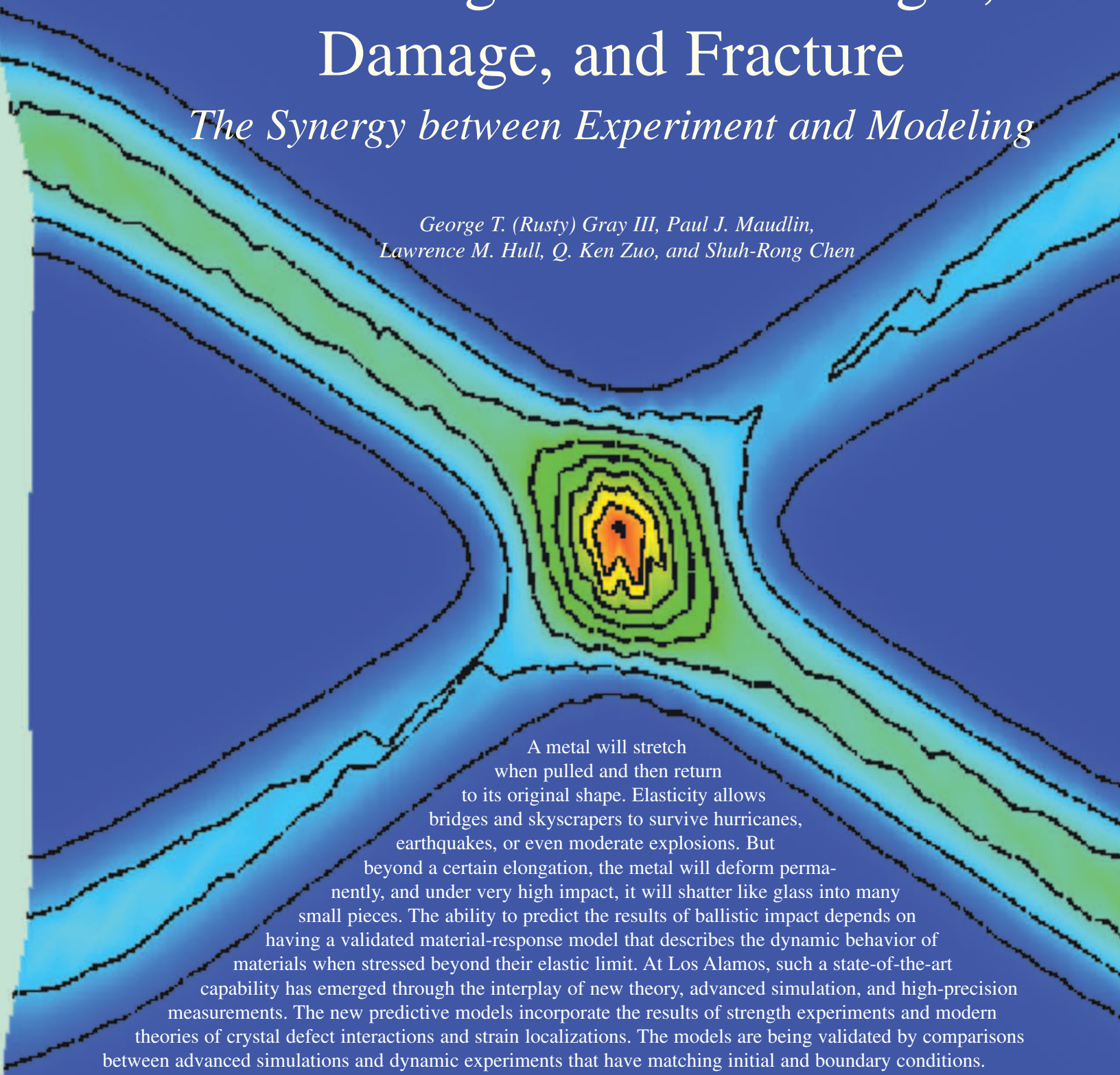
Saumon, D., and T. Guillot. 2004. Shock Compression of Deuterium and the Interiors of Jupiter and Saturn. *Astrophys. J.* **609** (2): 1170.

For further information, contact Lee A. Collins (505) 667-2100 (lac@lanl.gov).

Predicting Material Strength, Damage, and Fracture

The Synergy between Experiment and Modeling

George T. (Rusty) Gray III, Paul J. Maudlin, Lawrence M. Hull, Q. Ken Zuo, and Shuh-Rong Chen



A metal will stretch when pulled and then return to its original shape. Elasticity allows bridges and skyscrapers to survive hurricanes, earthquakes, or even moderate explosions. But beyond a certain elongation, the metal will deform permanently, and under very high impact, it will shatter like glass into many small pieces. The ability to predict the results of ballistic impact depends on having a validated material-response model that describes the dynamic behavior of materials when stressed beyond their elastic limit. At Los Alamos, such a state-of-the-art capability has emerged through the interplay of new theory, advanced simulation, and high-precision measurements. The new predictive models incorporate the results of strength experiments and modern theories of crystal defect interactions and strain localizations. The models are being validated by comparisons between advanced simulations and dynamic experiments that have matching initial and boundary conditions. Predictions of metal behavior—from strain hardening to fracture—show excellent agreement with the results of small-scale impact tests, tensile tests, and explosively driven fracture of metal shells.

From the first time one sharp object was used to shape another or cause another to fracture, the mechanical properties of materials—strength, ductility, and susceptibility to fracture—have shaped human history. Materials influence human life so profoundly that some have become synonymous with different eras—the Stone Age, the Bronze Age, the Iron Age, and the Nuclear Age. It is very possible that the current era, marked by people's growing dependency on electronics, may soon be dubbed the Silicon Age.

During the past eras, materials were selected almost exclusively on the basis of hands-on experience—one material shows favorable properties over another for a given application. But a new capability is now emerging—that of predicting material behavior and designing and engineering custom materials with predetermined characteristics. This trend could, in principle, lead to the age of "predictive materials technology," but only time will tell. What we demonstrate in this article is an emerging capability to predict and engineer the behavior of metals—their mechanical response under extreme loading conditions.

Engineering the response of metals and alloys to loading is an age-old trade, extending from the famous fifth century steels of Damascus to the aluminum alloys that enabled the modern era of civilian aviation. Manufacturing recipes were typically developed through trial and error, but during the years leading up to World War I, scientists and engineers conducted the first systematic studies and began to understand how the relationship between the applied stress, or force over area, and the resulting strain, or change in length, varied with temperature, strain rate, and stress state. That knowledge was quickly applied to critical wartime needs: high-speed manufacturing of metal parts (including high-speed wire drawing and cold-rolling of metal parts) and advances in ballistics, armor, and detonation physics. Spinoffs from those early studies led to increasingly sophisticated materials of relevance to defense, transportation, and communications.

In the last four decades, defense-oriented research has pushed the frontier of knowledge beyond standard stress-strain relationships to the complex mechanisms that occur under impact, namely, deformation, damage evolution, and fracture of metals and alloys. The basic mechanisms controlling those processes began to be understood, and the resulting models were used to estimate material response during high-speed impact, or high strain-rate, situations both natural and man-made. Familiar examples include automotive crash-worthiness; aerospace impacts, including foreign-object damage, such as that caused when a jet engine accidentally ingests a bird or a meteorite impacts a satellite; structural accelerations such as those occurring during an earthquake; high-rate manufacturing processes such as high-rate forging and machining; and conventional ordinance behavior and armor/antiarmor interactions. Within the past two decades, as computer power has grown and materials models have become more predictive, the R&D community has used, wherever possible, large-scale three-dimensional (3-D) computer simulations of these complex dynamic events in place of direct experimentation. The reasons are twofold: Either an experiment would be prohibitively expensive (a full-scale bird-ingestion test on a commercial jet engine conducted for the Federal Aviation Agency, for example, costs millions of dollars to field), or the system is too difficult to evaluate accurately through experiment (for example, the impact of a meteorite on the space station). In turn, the growing reliance on 3-D simulations of complex engineering systems has led to a growing demand for robust predictive material-response models.

We are developing predictive models for mechanical behavior through the interplay between systematic experiments and theory. In this article, we focus most heavily on the role of experiments in developing and validating mechanical response models. First, we present state-of-the-art experiments to measure very accurately the basic stress-strain relationships of metals and alloys under varying temperatures and strain rates. Second, we discuss the ways in which we measure and model specific damage evolution mechanisms that progress from the loss of load-carrying capacity to fracture. Finally, we present specialized experimental methods, measurements, and models describing the dynamic deformation and failure induced by explosive deformation. Although further experimental research and engineering work remain, the efforts described here demonstrate significant progress in quantifying the dynamic mechanical

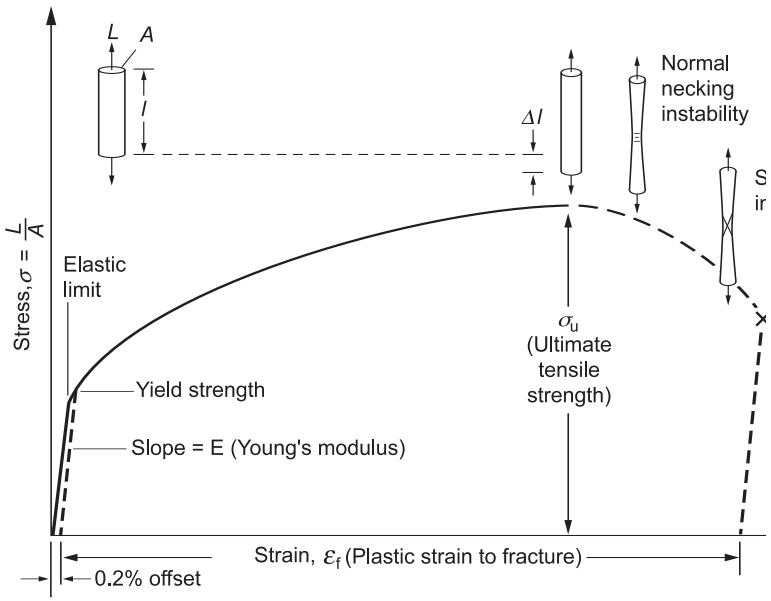


Figure 1. Tensile Stress-Strain Curve

The tensile test is the most common test used to measure mechanical properties. Round-bar or sheet samples are gripped at their ends and pulled at constant velocity (nominally, at constant strain rate) until they fail. Load and displacement of the sample are measured and plotted as stress σ (load/cross-sectional area) vs strain ε (sample elongation/original length). The elastic region, represented by Hooke's law ($\sigma = E\varepsilon$, where E is an elastic constant known as Young's modulus), is linear and reversible. The point of deviation from linearity is called the elastic limit and marks the onset of permanent deformation, or plastic strain. Because the onset of deviation is often very gradual, the "yield strength" of a metal is defined as the stress at 0.2% permanent (or plastic) strain. Continued plastic flow beyond the elastic limit produces increasing stress levels, a process called work hardening. During this stage, the sample deforms uniformly, elongating and thinning while the volume remains constant, until work hardening can no longer keep up with the continuing increase in stress caused by the reduction in the sample's cross-sectional area. At this point, the stress goes through a maximum, called the ultimate tensile strength, and the sample begins to deform nonuniformly, or neck, before it fractures in a ductile manner. Necking can reflect either "normal" or shear localization preceding fracture. In soft, annealed fcc metals, the typical total plastic (or permanent) strain immediately before fracture is 20% to 50%.

response of materials and applying those insights to the development of predictive material models of relevance to the defense mission of Los Alamos.

Mechanical Strength Models: Development and Validation

Standard mechanical strength models for metallic materials spell out the relationship between stress (load per unit area of material) and the resulting strain (change in length, area, or volume relative to the original dimension) during elastic and stable plastic deformation (see the positive-slope side of the stress-strain curve in Figure 1). However, at some background strain, metals will transition from uniform, or homogeneous, deformation to heterogeneous, or localized, unstable behavior (occurring usually on the negative-slope side of the stress-strain curve). In fact, when an as-received material is pulled at a constant velocity at the boundaries, it follows a stress-strain curve comparable to that in Figure 1. This stress-strain path has four distinct stages: (1) uniform, or homogeneous, deformation and accumulation of background strain, (2) material instability or bifurcation (a condition that indicates loss of load-bearing capacity), (3) transition to heterogeneous, or localized, deformation (in a normal and/or shear mode), and (4) accumulation of damage (small cracks and voids) that ultimately coalesces into a fracture surface. In the description of the experiments and modeling provided in the sections below, both homogeneous and localized deformations are investigated and modeled at macroscopic scales. To be predictive, those models must capture the fundamental relationships connecting the independent variables of stress, strain rate, strain, and temperature to specific bulk material responses such as yield stress or flow stress, strain hardening, texture evolution, evolution of global damage, subsequent heterogeneous damage, such as strain localization and cracking, and finally, material failure. Moreover, for the applications of interest, we need to predict those responses accurately for such extreme conditions as large deformation and high strain rates, pressures, and temperatures. Our materials models must therefore be based on quantifiable physical mechanisms, characterized with inexpensive direct experiments, and validated through comparisons with results of small-scale and integral tests.

Standard Measurements of Strength. One develops a strength model for a particular material by measuring its mechanical properties. The samples of interest are loaded in

compression, tension, or torsion over a range of loading rates and temperatures germane to the application of interest. Various mechanical testing frames are available that achieve nominally constant loading rates for limited plastic strains and, thereby, a constant strain rate. The standard screw-driven or servo-hydraulic testing machines achieve strain rates of up to 5 per second. Specially designed testing machines, typically equipped with high-capacity servo-hydraulic valves and high-speed control and data acquisition instrumentation, can achieve strain rates as high as 200 per second during compression loading. To go even higher, we must employ projectile-driven impacts that induce stress-wave propagation in the sample materials. Chief among these dynamic loading techniques is the split-Hopkinson pressure bar (SHPB) (Gray 2000), which can achieve the highest uniform uniaxial compressive stress loading of a specimen at a nominally constant strain rate of about 10^3 per second. In fact, we readily reach strain rates of up to 2×10^4 per second and true strains of 0.3 in a single test by using the SHPB. At these stresses and strain rates, however, the uniformity of stress loading and constancy of strain rate are not guaranteed, so care must be exercised.

Developing a Strength Model for Tantalum. As shown in Figure 2, we used several of the testing techniques just mentioned to measure the stress-strain response of unalloyed tantalum metal over a wide range of strain rates and temperatures. We then determined the unknown parameters in our strength model by fitting the model to the experimental data. In general, this model predicts flow stress (level of stress needed to produce dislocations in the crystal lattice as a function of strain, strain rate, and temperature), and the mathematical form used is known as the Mechanical Threshold

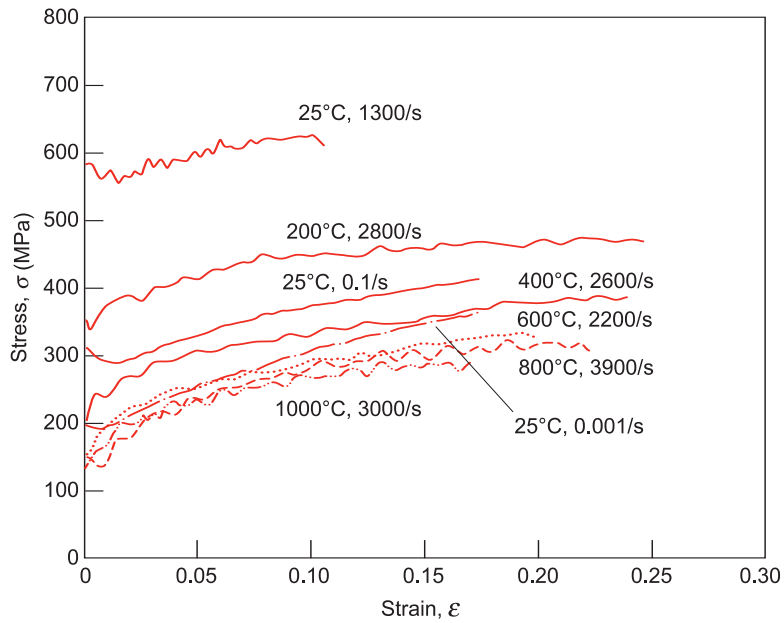


Figure 2. Compressive Stress-Strain Curves of Unalloyed Tantalum
Under lower applied stress, the material deforms more readily as the temperature is increased; conversely, the stress required to deform the material increases as the strain rate (rate of applying stress) is increased.

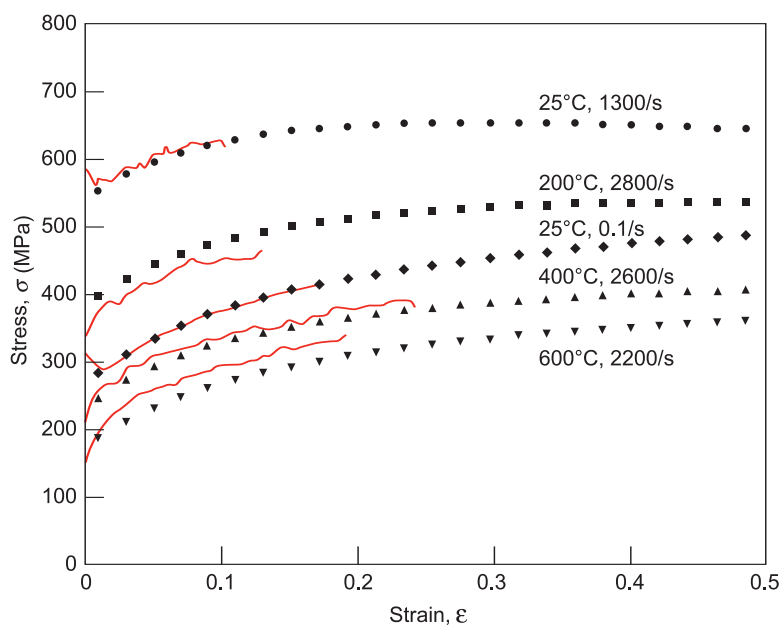


Figure 3. Mechanical Behavior of Stress-Strain Curves for Unalloyed Tantalum: Experiment vs Calculation
The experimental mechanical behavior of the stress-strain curves (red lines) of unalloyed tantalum measured for a range of temperatures and loading rates are compared with the fit to the MTS model. Experiment and calculation agree very well.

Strength (MTS) Model (Chen and Gray 1996). Figure 3 gives an example of the characterization of the model parameters for unalloyed tantalum data that show the model accurately capturing the dependence of yielding and strain hardening as strain rate and temperature are varied.

A key feature of the MTS model is an internal state-structure variable that describes the physical property of work hardening that occurs as a metallic specimen deforms plastically. This hardening variable evolves in the context of the model as mobile dislocations (lattice defects), created during the deformation process, interact with other stored-dislocation structures. Those in turn evolve via the dynamic microscale processes of mobile-dislocation storage and stored-dislocation annihilation, both controlled by the independent variables mentioned above. This scalar representation of dislocation behavior by the MTS model is fairly accurate for a large number of metals used in predictive engineering simulation. The model is also easily extended if one uses the concepts of plastic potential and yield surfaces, physically based on micromechanics of polycrystal plasticity, to describe directional (anisotropic) plastic deformation (Maudlin et al. 1999). This extension of the model is illustrated below.

The problem is that many engineering problems, such as foreign-object damage and ballistic impact, involve strain rates of 10^4 to 10^5 per second, values that are well beyond the range accessible for direct measurement. Is our strength model valid at those higher strain rates? Since we cannot test the model directly, we use it to predict the results of a simple validation test such as the Taylor cylinder impact test. This readily conducted axisymmetric test realizes strain rates as high as 10^5 per second and deformations in excess of 100 percent. The MTS model is used as part of the constitutive module in a 3-D finite-element continuum mechanics code for simulating the Taylor test, and the results are then compared with post-test geometries (for example, the cylinder side profiles) for several impact velocities.

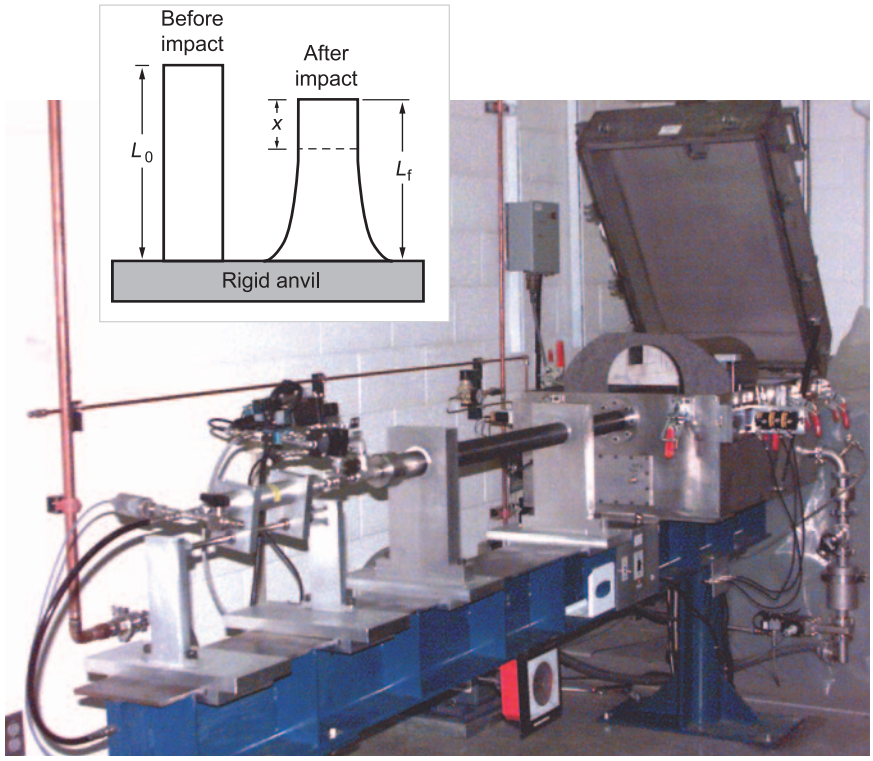


Figure 4. The Los Alamos Taylor Impact Test Facility
The photograph shows the apparatus used to fire a small cylindrical test sample at high velocity against a massive, rigid target, and the inset shows the initial and final states of a cylindrical sample.

The Taylor Impact Test for Validating the Strength Model

The Taylor cylinder impact test shown in Figure 4 was developed during World War II by G. I. Taylor (1948) to screen materials for use in ballistic applications. It entails firing a small solid cylinder rod of some material of interest, typically 7.5 to 12.5 millimeters in diameter by 25 to 40 millimeters in length, at high velocity against a massive and plastically rigid target. As indicated schematically in the inset to Figure 4, the impact plastically deforms and thereby shortens the Taylor rod by causing material at the impact surface to flow radially outward relative to the rod axis. By assuming simple one-dimensional plastic flow, Taylor related the fractional change in the rod length (difference between the final length L_f and the initial length L_0) to the flow stress, one point on the stress-strain curve in Figure 1.

The Taylor impact test represents an escalation of complexity relative to tests made with the split-Hopkinson pressure bar. Rather than measuring the stress-strain response at a uniform stress state and strain rate, a Taylor test involves gradients of stress, strain, and strain rate integrated over time to produce a final strain distribution. In fact, the Taylor test is most often used to intentionally probe the deformation responses of metals and alloys in the

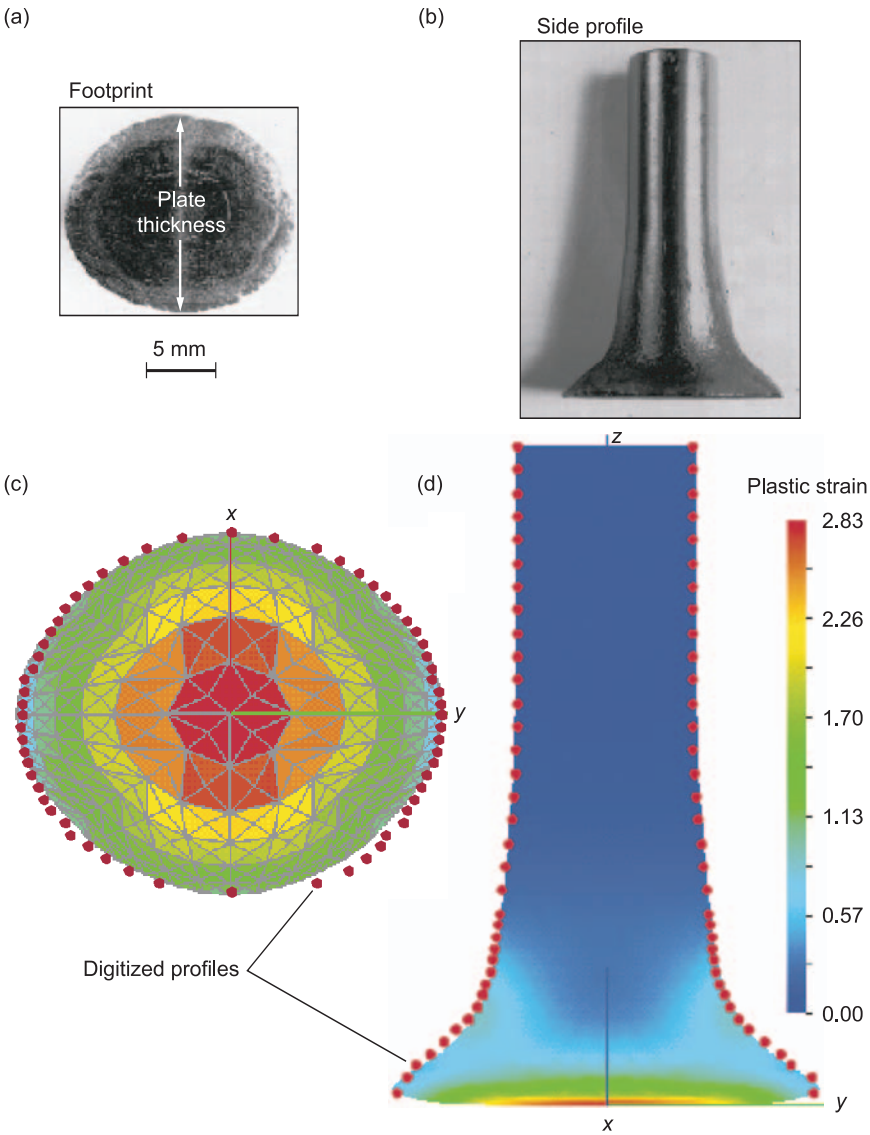


Figure 5. Post-Test Geometry of Taylor Specimen and Simulation Results

Photographs of the post-test geometry for a tantalum Taylor specimen show (a) the footprint and (b) the side profile. The colored patterns in (c) and (d) represent the plastic strain distribution predicted by the EPIC 3-D code simulation for the footprint and major side profile, respectively, of the Taylor sample. The red dots are the digitized experimental profiles.

presence of large gradients of stress, strain, and strain rate. Nevertheless, 3-D finite-element simulations of this integrated test have proved to be highly sensitive to the accuracy of material strength models used in the numerical codes. From a comparison of the cylinder profile of the post-test Taylor sample with the profile predicted by the finite-element code simulations, we have determined how well the material model and the code implementation describe the spatial gradients of deformation stress and the strain rates that ultimately lead to the final strain distribution seen at the end of the test.

The spatial stress and strain gradients and the strain rates, in turn, are direct measures of the stress-strain tensor described in the strength model. Moreover, based on tests conducted on copper, tantalum, aluminum, tungsten-nickel-iron alloy, and steels over a range of impact velocities, we have seen that the final strain distribution is sensitive not only to strain hardening, strain rate, and temperature, but also to crystallographic texture.

Figure 5 compares experimental and finite-element simulation results for a Taylor test specimen (Maudlin et al. 1999) of unalloyed tantalum that has a moderately strong crystallographic texture (that is, a directional dependency of its flow stress due to material processing). The cylinder had been initially cut from a rolled tantalum plate with a preferred texture (crystal orientations) associated with the rolling process. In this particular process, the {111} planes (that is, the major diagonal planes) of individual cubic crystals were most often stacked with the normals to the {111} planes aligned with the through-thickness direction of the plate, which represents a strong direction in this particular stock of tantalum. Because the Taylor cylinder was cut with its axis perpendicular to the through-thickness direction of the plate, the material strong direction is aligned perpendicular to the loading axis of the impact; thus, one transverse direction of the initially round Taylor rod is stronger than the other. Consequently, the impact and subsequent plastic deformation during the Taylor test produced an anisotropic mechanical response illustrated by the elliptical footprint and the side profile shown in Figures 5(a) and 5(b). After testing, we used an optical comparator to generate a digitized footprint of the cross-sectional area at the impact interface and digitized side profiles—see the red dots in Figures 5(c) and 5(d). We simulated the Taylor impact test with

the explicit, Lagrangian, finite-element code EPIC in a 3-D mode by using the MTS model for tantalum. The cylinder was spatially modeled using 4185 nodes and 17,280 single-integration-point tetrahedral elements. Because in the experiment both the anvil and cylinder base had mirror-like finishes and were carefully aligned for orthogonal impact, interfacial friction at impact at the cylinder-anvil interface was negligible and could be ignored in the simulation. Similarly, the vanishing hardness of the Taylor cylinder relative to that of the anvil precluded any measurable plastic compliance within the anvil, and so that too could be ignored. We simulated the impact event for 90 microseconds of problem time, after which plastic deformation reached quiescence, as it had in the experiment. Calculational results of an impact-interface footprint and a late-time cylindrical major profile are shown in Figures 5(c) and 5(d). These results are compared with the experimental shapes indicated by red dots. The calculated elliptical footprint shown in Figure 5(c) has an eccentricity (ratio of major to minor diameters) of about 1.20 that compares well with the experimental footprints. If the sample were to have been isotropic, in which case the individual crystals would have been randomly oriented, it would have produced a round footprint with an eccentricity of 1. The major side profile compared with experimental data (red dots) in Figure 5(d) indicate that the final length agrees well with the experimental length and that the axial distribution of plastic strain also tracks very well with the experimental profile.

The MTS model was used together with an anisotropic yield surface whose tensor implementation of texture allows the modeling of crystalline texture. The combination produced very good agreement between the calculated and experimental plastic deformation field for the tantalum cylinder, including the anisotropic shape of the final cylinder. This example demonstrates the state of the art of 3-D mechanical modeling of yield anisotropy for a material subjected to a fairly complicated impact test, in which spatial variations in stress, strain, and strain rate occur simultaneously. Further work on materials demonstrating increasingly complicated deformation mechanisms as part of their mechanical stress-strain behavior will determine the direction of future mechanical-behavior model development.

Validating Strain-Localization and Fracture Models

The Taylor test is a good example of bulk deformation in which spatial gradients of permanent plastic strain extend over the entire sample—from 100 percent strain at the impact interface to near-zero strain at the opposite end of the cylinder. Under more extreme loading conditions, however, material instabilities can set in, causing plastic deformation to localize into planes that extend through the metal. Those localized regions of strain often appear as bands in the post-test specimen sections. Such material instability and plastic localization lead to the loss of load-carrying capacity and to damage evolution and fracture, as depicted on the right side (negative-slope region) of the stress-strain curve in Figure 1. In a typical tensile experiment, a material is pulled at constant strain rate, and its load-carrying capacity first increases (strain hardening). At some point, however, the load-carrying capacity reaches a maximum, and if the applied load remains constant or increases while the load-carrying capacity begins to decrease, a run-

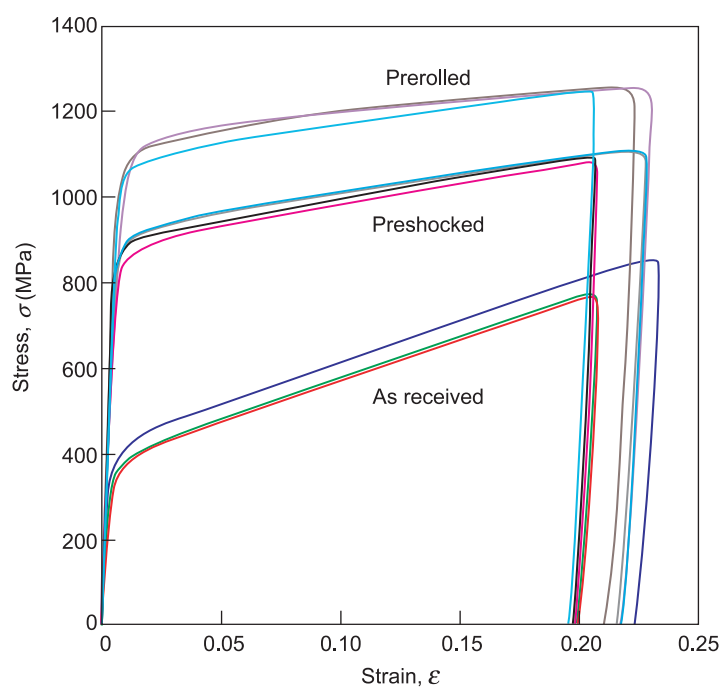


Figure 6. Uniaxial Stress Compression Measurements
The curves are of experimental stress vs strain for a stainless steel (SS316L) rolled-plate stock material. Shown are room temperature, quasistatic, and uniaxial-stress compression results as measured for the three SS316L conditions.

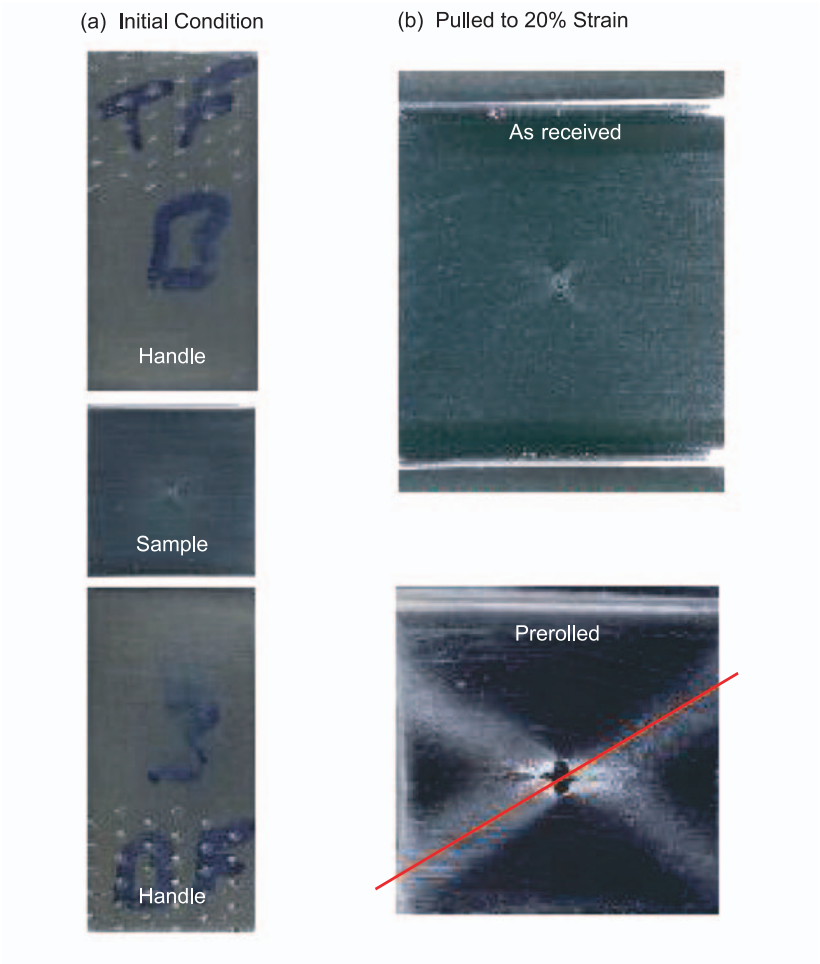


Figure 7. Stainless Steel Flat-Plate Shear Specimens
(a) The photograph is of a flat-plate SS316L sample of uniform-thickness rolled stock before the tensile test. The initial specimen geometry includes a small defect (hole) in the center of the sample. (b) The photographs are of two specimens after having been stretched quasistatically at room temperature to a 20% strain. The initial material states were as received (top) and prerolled (bottom).

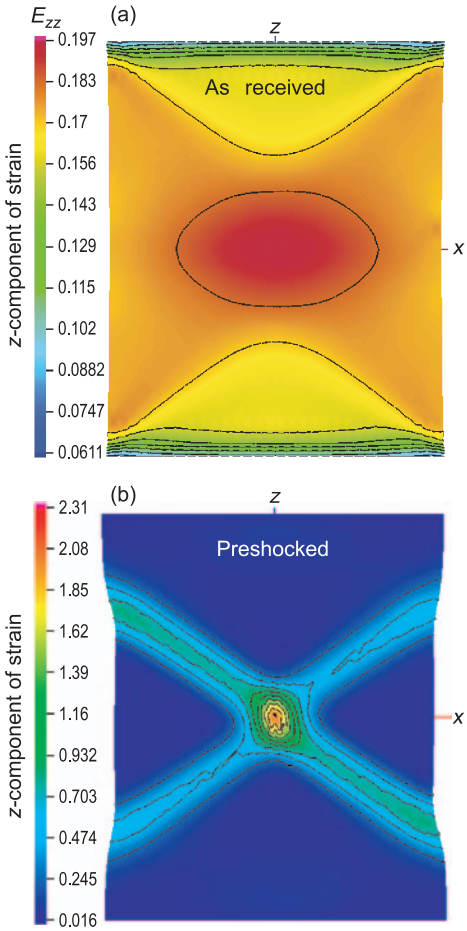


Figure 8. Strain Distributions for Flat-Plate Shear Specimens
Results of a 3-D finite-element simulation are shown in terms of the Lagrangian strain E_{zz} at 100 μs into the deformation. Simulations use the MTS model for the stainless steel sample with an initial hole defect. Shown in (a) is the final strain state in the as-received material; in (b), deformation localizes into a shear band doublet in preshocked material.

away or unstable situation will occur with increasing strain. In many materials, this unstable motion results in localized deformation.

We have used a number of small-scale experiments, including pulling simple flat-plate samples, to study localized-strain and damage evolution as a function of the "starting state" of a material. The starting state could be a prerolled or preshocked process with an amount of flow stress hardening (slope of Figure 1) that depends on the magnitude and direction of strain applied during preprocessing. In these experiments, small, uniaxial, square flat-plate tensile specimens (12.7 millimeters on one side and 1.0 millimeter in thickness) are cut from a stainless steel (SS316L) rolled-plate stock, either as received or preprocessed, and the specimens are then pulled to failure and insipient failure (about 40 percent longitudinal true strain).

Before the flat-plate testing, we measured the mechanical properties of the starting material by cutting cylindrical samples from the same preprocessed stock and subjecting them to uniaxial compression testing. Figure 6 compares flow stress curves from these tests for three material starting conditions: an as-received fairly ductile material, the same as-received material but prerolled by a 20 percent strain before testing, and the same as-received material but preshocked by a high-

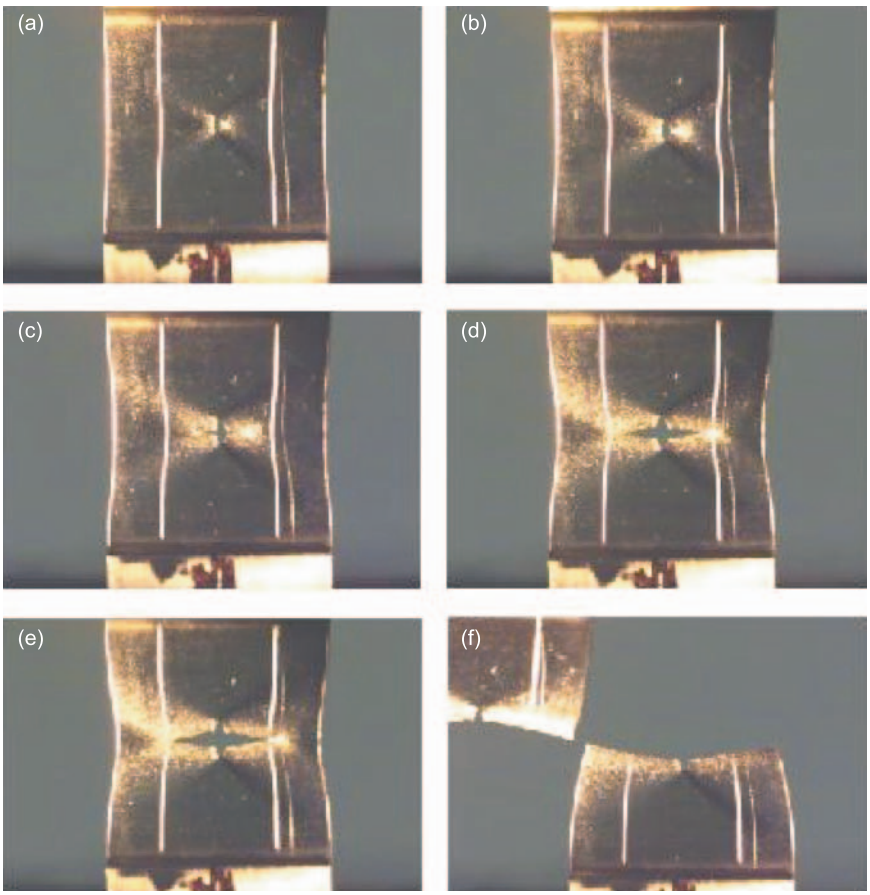


Figure 9. SS316L Flat-Plate Shear Specimen—Late Deformation and Fracture Stages

Dynamic photographs of an SS316L flat-plate shear specimen show it during the late stages of deformation and fracture. This test was conducted quasi-statically at room temperature for the initial, prerolled, hardened material condition. The initial specimen geometry has a small defect (hole) in the center of the gauge section. Time progresses from (a) to (f).

explosive (PBX-9501) plane detonation wave loading. The preshocked material has a higher flow stress and a lower slope than the as-received material. It is therefore less stable or closer to the point at which material instability initiates a transition from homogeneous to localized deformation. This reduction of stability in preshocked material is often relevant to defense applications. Figure 6 also shows that the prerolled material has even higher flow stress and lower slope, and therefore less stability, than the preshocked material.

The photographs in Figure 7(a) show the geometry of the flat-plate tensile test, including the handles used to pull the sample. All the flat-plate test specimens have a small initial mechanical defect (hole) that has been machined into the center of the sample. Figure 7(b) shows magnified photographs of the as-received and the prerolled specimens after having been pulled to about 20 percent strain. These specimens exhibit post-yield (the strain exceeds the elastic limit) shear bands. In particular, each exhibits a strain localization doublet centered on the hole. The prerolled material is obviously much more unstable. Compared with the as-received specimen, it exhibits a very prominent strain localization doublet. This higher susceptibility to instability and strain localization can be anticipated from the stress-strain results of Figure 6, in which the flow stress during plastic deformation is higher in magnitude and lower in slope for the prerolled specimen. The orientation of the localization doublet in the prerolled sample measured relative to a transverse specimen direction (a horizontal axis) is $\beta \cong 30° \pm 1°$.

In 1975, Rudnicki and Rice achieved a theoretical breakthrough by formulating a mathematical description of strain localization that treats the jump in material strain as a stationary wave discontinuity, formally analogous to the description of a shock wave as a traveling wave discontinuity. In many dynamic applications, localizations of strain lead to material damage (voids and cracks) and final failure of system components. The localization description developed by Rudnicki and Rice is a cornerstone in the material instability or bifurcation literature. The theory, which becomes applicable when the strength of a material becomes saturated (the stress reaches a maximum, and the slope of the stress-strain curve is zero or negative), predicts three items of interest: the onset of material instability, the orientation of the localization planes, and the direction of the straining jump in the localization band. Figure 8 shows our finite-element predictions for the SS316L as-received and preshocked samples. The simulations validate the fundamentals of the Rudnicki and Rice derivation, predict the background strain before the onset of localization, and predict the orientation of the localization planes. Deformation appears as mostly uniform, or homogeneous, in the as-received specimen—Figure 8(a)—in contrast with the preshocked speci-

men—Figure 8(b)—where it localizes into a doublet of shear bands. These simulations are then compared with experiments to validate the models. The predicted localization orientations are in good agreement with the test results. Despite the shear doublet apparent in the prerolled test specimen shown in Figure 7, all specimens in all three starting material conditions, with or without an initial defect, failed in a normal mode; that is, they fractured transversely across the specimen. The dynamic sequence of photos in Figure 9 shows the development of this horizontal fracture and reveals geometric specimen necking just before fracture. The latter implies a value of stress triaxiality (that is, ratio of pressure to flow stress) larger in magnitude than the uniaxial value of 1/3. As confirmed by additional bifurcation analyses with the Rudnicki and Rice theory, this transition from uniaxial to higher-stress triaxiality, which causes more lateral restraint, is responsible for rotating the strain-localization planes and producing the appearance of a horizontal fracture; the photos show a fracture edge in the observation plane, where a macro crack runs in the horizontal direction across the section intersecting the hole. This localization and fracture phenomenology manifests itself in the context of explosive loading problems involving more complex states of stress, as will be discussed in the next section.

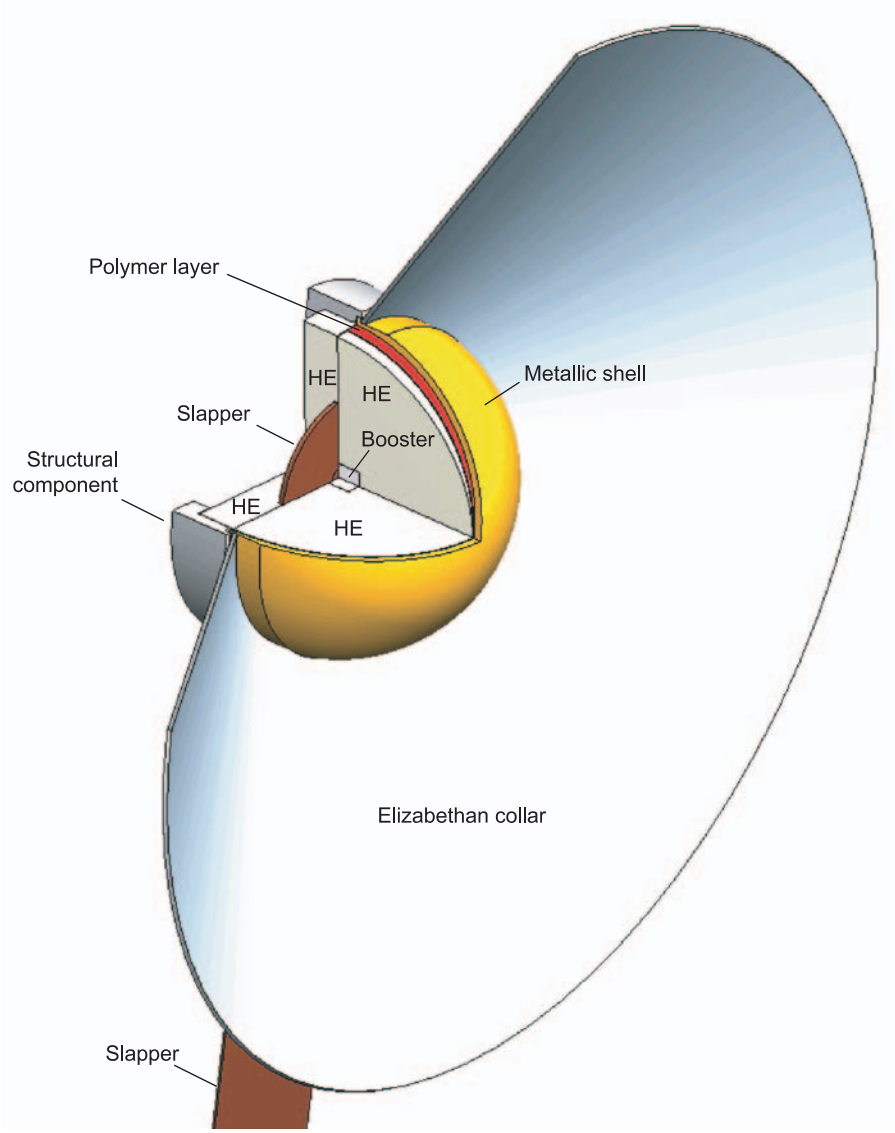


Figure 10. Solid Model of the Filled-Hemishell Experimental Hardware
The experiment conducted with this hardware was designed to reproduce as closely as possible a spherical detonation inside a metal hemishell.

Validating Models for Explosively Driven Dynamic Deformation

Many defense applications at Los Alamos involve explosively driven systems in which the materials are subjected to extreme conditions of temperature and strain rate. Our ultimate goal for modeling and simulation is to develop the ability to predict the onset of strain localization in shells of arbitrary geometry, the coalescence of those localizations into a network of cracks as a precursor of fracture, the fracture of the shell into individual fragments, and the size, velocity, and spatial distribution of those fragments. We must be able to model and simulate the correct physics for a broad variety of materials that are manufactured into a shell geometry and tested under various loading configurations. In this section, we describe small-scale integrated tests involving high explosives that are exploited as another source of validation data (albeit, more complicated and challenging than, for example, the flat-plate tensile specimens above) for modeling and simulation. One such test is the explosively driven hemispherical shell. In the simulation of this experiment, a hemispherical metal shell filled with explosive is initiated at the spherical center. Radial propagation of the spherical detonation

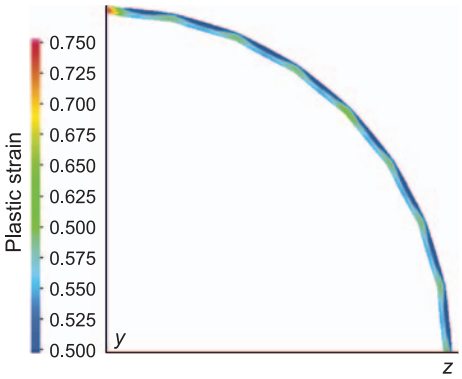


Figure 11. Predicted Strain Distribution of the Filled Hemishell

This calculation used as-received material strength properties. The plastic strain is the plotted field variable. The results show fairly uniform deformation.

wave results in the simultaneous arrival of the wave at the explosive-metal interface.

Experimental Design. A hemisphere is filled with high explosive (PBX-9501) whose density is 1.833 grams per cubic centimeter (g/cm^3). Both the explosive and metal are highly characterized in terms of mechanical properties and process control. Also, no effort was spared to make the fielded design as close to the idealized geometric configuration as possible and to facilitate a clear view of the fracture process (see Figure 10). For example, the high-explosive slapper system for initiating the detonation was designed to approximate as closely as possible the mathematical idealization of a detonation initiated at a single point. The booster pellet is embedded in the main charge so that the initiation system should not perturb the explosive drive at the pole. There is also no metal case around the booster, and the slapper itself is a thin copper/plastic laminate system that generates a minimal amount of debris that could contaminate the optical view of the fracture process. The fixture that holds the hemishell configuration in space is a thin (0.75-millimeter) "Elizabethan collar" made from spun aluminum. The collar provides a lightweight but robust symmetrical mount for the shot. A 5-millimeter-thick disk of explosive holds the initiation assembly in place, and a polycarbonate ring holds the explosive, collar, and metal shell in place. As the shell expands, the explosive interacts with the collar and pushes it out of the region of interest. The collar also disperses the debris from the slapper initiator and increases the late push on the hemishell's equator. These design elements cause the motion of the shell to approximate more closely a spherical expansion.

The filled-hemishell design is more attractive for modeling and simulation validation activities than the classic design of end-detonated filled-cylinder devices. The filled hemishell prevents the seeding of strain localizations that occur in filled cylinders when the detonation wave sweeps from one end to the other along the explosive/metal interface. The spherical symmetry of the detonation in the filled hemishell thus allows direct observation of strain localization, controlled by the material instability and bifurcation concepts discussed above. The resulting nearly spherical expansion of the shell causes predominantly biaxial stress, resulting in fragments with aspect ratios near unity, and maintains axis symmetry so that the 3-D effects observed in the shell fragmentation process can be linked to the fracture process itself. Finally, because the test configuration is small, various diagnostics, including proton radiography, become quite feasible for this integrated test.

Simulation Results for a U6Nb Hemishell. Three-dimensional finite-element simulations of the filled hemispherical shell of uranium alloyed with 6 percent niobium were conducted with the material modeling described above. U6Nb is a good test of our modeling capability because tensile tests, such as the flat-plate tensile tests described above, have shown preshocked U6Nb to be more unstable with a strong propensity to strain-localize when compared with most metals. Figure 11 shows predictions of the plastic strain distribution in a centerline section of the U6Nb shell for a stable, as-received material condition, showing fairly uniform deformation (plastic strain ranging from 0.5 to 0.75); these results correspond to a time of 5 microseconds after the detonation wave loads the shell. The computations were performed with the EPIC finite-element code with the goal of predicting the onset of material instability in the U6Nb material. The 3-D shell shown in Figure 12 shows pronounced strain localization (the plastic strain reaches 1.0 in several cells, whereas most of the hemishell has a strain of about 0.6), resulting from the unstable material character of shock-processed U6Nb. The shell expands axisymmetrically at small times, while the material is still stable, but quickly loses this symmetry and exhibits 3-D effects as strain localizations develop.

The simulation results are consistent with our understanding of material instability and strain localization; the critical strain at which the material loses stability increases as the hardening modulus (the slope on the stress-strain curve of Figure 1) increases, and it decreases as the magnitude of the flow stress increases. The main effects of preshock on U6Nb are to raise the initial flow stress and to significantly reduce the hardening modulus. Consequently, the shock-hardened U6Nb becomes unstable shortly after the detonation wave loads the shell, and subsequently the strain localizes, and the shell expands nonuniformly.

Experimental vs Simulation Results. The EPIC predictions were compared with the experimental results from the corresponding filled-hemishell test. Both proton radiographs taken during shell expansion and the fragments recovered following the shots yield experimental information on the onset of strain localization. Figure 13 presents proton radiographs viewed normal to the pole of the hemishell at different times. Localized thinning is apparent at an early time (8.2 microseconds)—Figure 13(b). At a later time, these localizations coalesce into an ultimate fragmentation pattern—Figure 13(c). Supporting fragment recovery experiments that use a water medium have yielded significant information on the failure of the filled hemishell. In these recovery experiments, the shell is initially immersed into water so that there is no metal/water impact that could induce additional material damage. The recovered fragments represent an approximation to the conditions at the time at which fragmentation is complete (the fragments are fully separated from each other, as shown in the 16.8-microsecond radiograph of Figure 13).

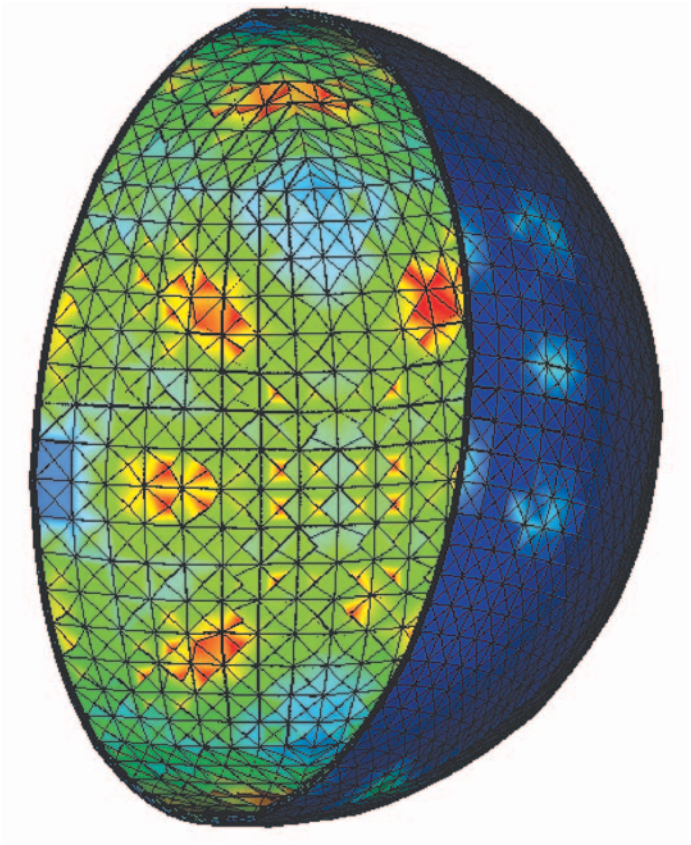


Figure 12. Predicted Strain Distribution of the Preshocked Filled Hemishell
For this calculation, we used shocked material strength properties. Plastic strain is the field variable indicated as color contours. Development of pronounced strain localization is apparent.

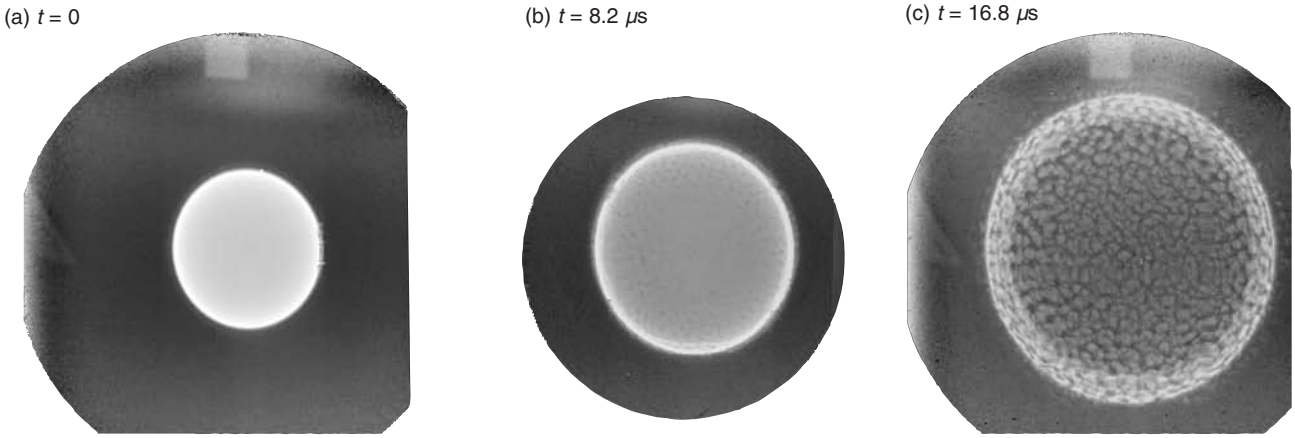


Figure 13. Unstable Expansion of a Filled Hemishell Shown by Proton Radiographs
The proton radiographs taken during shell expansion yield experimental information on the onset of strain localization. Localized thinning is apparent, which then develops into a fragmentation pattern at 16.8 μs.

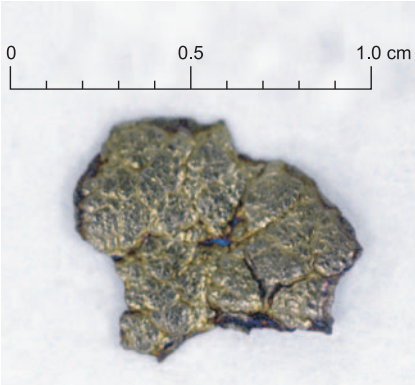


Figure 14. Photograph of a U6Nb Fragment
This photograph of a fragment recovered with a water medium reveals a free surface with many strain-localized features.

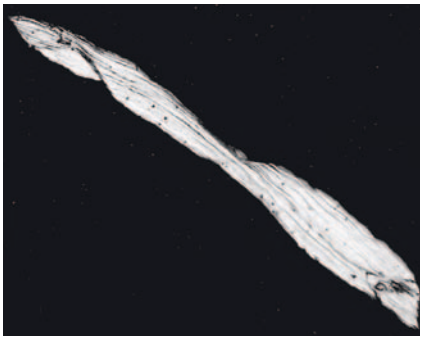


Figure 15. Edge Micrograph of a U6Nb Fragment
This is a microscopic side view of the fragment shown in Figure 14.

A surface photograph of a recovered fragment (Figure 14) reveals a free surface sculptured by many strain-localized features. Evidently, some strain localizations dominate and coalesce to form the fragments, while the growth of others is arrested as a natural consequence of the competition among localizations to accommodate the loading and boundary conditions. Figure 15 shows a microscopic side view of the fragment, obtained by a cut on a plane normal to the free surface. In regions distant from localizations, the background strains are relatively large at about 60 percent. Interestingly, the defects (for example, grain boundaries, niobium concentration bands, and carbide inclusions) that one might expect to influence the strain localization distribution appear not to be responsible for the initiation, or nucleation, of the localization phenomenology.

The experimental data support the computations qualitatively and, to some degree, quantitatively. At 8 microseconds, strain localization is evident in the experiment as thinning occurs in small areas. The recovered fragments show background strains of about 60 percent, and this value compares well with the predicted background strain at 8 microseconds, provided we assume that, when localization begins, the background strain ceases to increase and all the subsequent strain is concentrated in the localizations. The validity of this assumption remains to be checked. Also, since the physical basis for the nucleation of strain localization has not been identified in the simulations, we only expect the predicted distribution to roughly match the experimentally determined spatial distribution of thinned areas.

In summary, the shell experiments have validated our modeling and simulation capability to predict when a metal will bifurcate into localized strain, provided we have an accurate mechanical representation at the relevant conditions.

Conclusions

The Taylor cylinder impact test, the plane-strain tensile test, and explosively driven hemisphere test represent readily conducted experiments that probe the deformation, damage evolution, and fracture behavior of materials. Because these tests are very sensitive to large gradients of stress, strain, strain rate, and shock loading, we are using them to evaluate and validate the correctness of our mechanical models that are implemented and destined to be implemented into large-scale 3-D simulation codes.

Robust models that capture the physics of high-rate material response are required for developing predictive capability for highly dynamic events. The increased effort to link experiments and modeling within the computational mechanics community and the increased emphasis on code verification and validation within the Los Alamos National Laboratory defense programs are accelerating this development. These efforts are already receiving recognition through the recent establishment of verification and validation committees within various technical societies. ■

Further Reading

Chen, S. R., and G. T. Gray III. 1996. Constitutive Behavior of Tantalum and Tantalum-Tungsten Alloys. *Metall. Mater. Trans. A* **27** (10): 2994.

Gray III, G. T. 2000. Classic Split-Hopkinson Pressure Bar Technique. In *ASM–Handbook,* Vol. 8, Mechanical Testing and Evaluation, p. 462. Edited by H. Kuhn and D. Medlin. Metals Park, Ohio: ASM International.

Maudlin, P. J., J. F. Bingert, J. W. House, and S. R. Chen. 1999. On the Modeling of the Taylor Cylinder Impact Test for Orthotropic Textured Materials: Experiments and Simulations. *Int. J. Plasticity* **15**: 139.

Rudnicki, J. W., and J. R. Rice. 1975. Conditions for the Localization of Deformation in Pressure-Sensitive Dilatant Materials. *J. Mech. Phys. Solids* **23**: 371.

Taylor, G. I. 1948. The Use of Flat-Ended Projectiles for Determining Dynamic Yield Stress. I. Theoretical Considerations. *Proc. R. Soc. London, Ser. A* **194** (1038): 289.

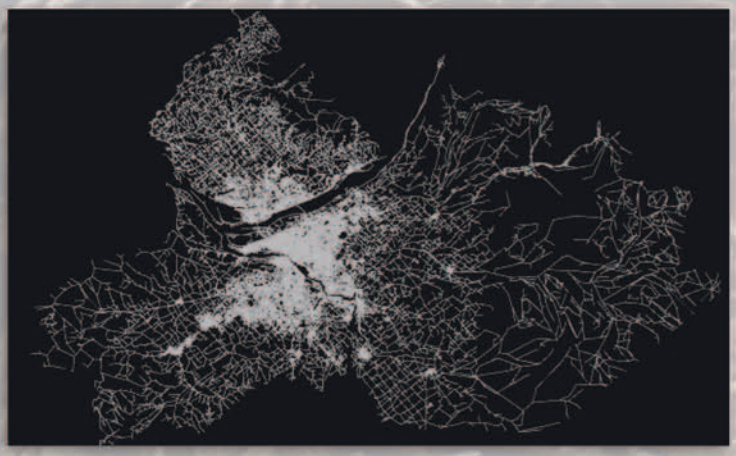
For further information, contact George T. Gray III (505) 667 5452 (rusty@lanl.gov).

Complex Networks

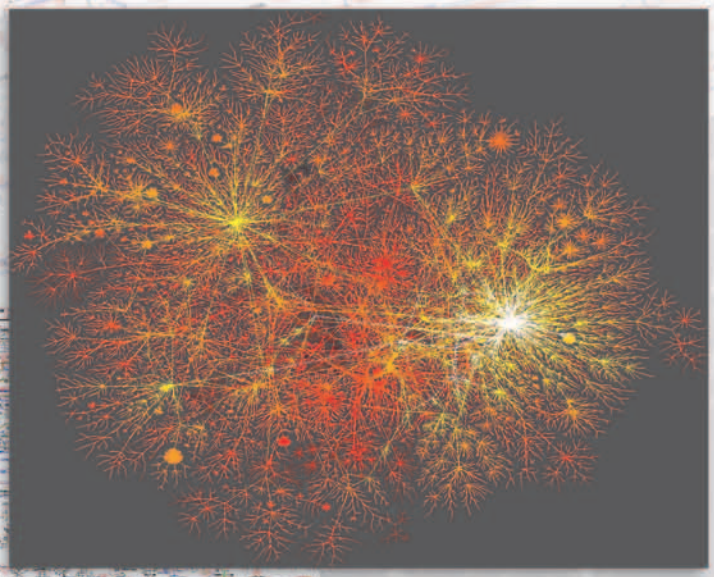
The Challenge of Interaction Topology

Zoltán Toroczkai

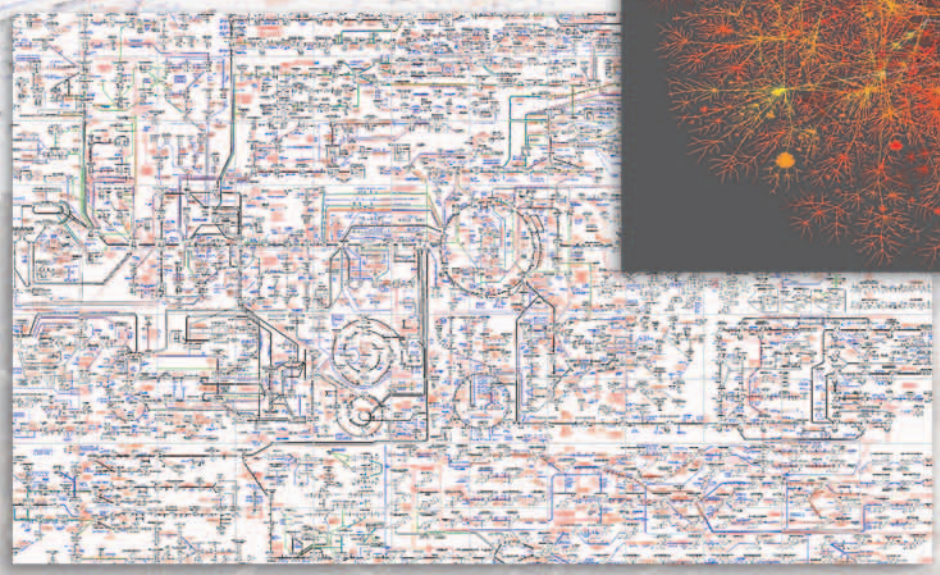
Networks have recently become a paradigmatic way of representing complex systems in which the pattern of interactions between a system's constituent parts is itself complex and is evolving together with the system's dynamics. Transport is the main function of these dynamic networks. It is therefore crucial that we understand the coupling between the network structure and the efficiency and robustness of the transport processes on the structure. Such understanding will have a huge impact, allowing us to control signaling processes in the cell and to design robust information and energy-transmission infrastructures, such as the Internet or the power grid. However, achieving this type of understanding is rather challenging, because of the discrete and random nature of network topology. This article reports on some of our results that connect network dynamics and transport efficiency. It also illustrates the power behind the ability to control the topology of the interactions in the design of scalable computer networks.



The roadways of Portland, Oregon.



Macroscopic snapshot of Internet connectivity (skitter data) with selected backbone Internet service providers. (This photo is courtesy of the Lumeta Corporation.)



The metabolic pathway. (Gerhard Michal: Biochemical Pathways,1999 © Elsevier GmbH, Spektrum Akademischer Verlag, Heidelberg.)

Systems of many interacting particles typically exhibit complex behavior. In most well-known complex systems, the topology of the interactions between particles can be described by simple structures, such as regular crystalline lattices or a continuum, and the complex behavior arises from nonlinearity and nonlocality, which describe the nature of the interactions themselves. There is, however, a large class of systems called complex networks, in which the interactions are mediated not by a continuum (or a simple regular structure) but by a complex graph, whose structure may evolve as part of the dynamics of the interactions.

Familiar systems in almost every area of life form such complex networks. Here are a few examples: transport and transportation infrastructures (electric power grids, waterways, natural gas pipelines, roadways, airlines, and others) social interactions (acquaintance networks, scientific collaboration networks, terrorist networks, sex webs, and others), communications networks (the World Wide Web, the Internet, microwave backbone, and telephone networks), biological networks (metabolic networks, gene regulatory networks, protein interaction networks), and networks in ecology (food webs). Although these systems have been known for a while, their complexity has been explored only recently because the large databases and the immense computational power required to analyze network data were almost nonexistent two decades ago.

Even a cursory "look" at the structure of real-world networks creates a breathtaking impression: These are large objects containing thousands, or sometimes, even hundreds of millions, of nodes with an intricate mesh of connections among them. For the last decade, the science of complex networks has focused on describing the structural complexity of real-world network topologies. By looking at the three images on these opening pages, one can easily surmise, that statistical and probabilistic methods are essential to that description. Today, the focus has expanded beyond network structure to an understanding of the relationship between structure and dynamics and the implications of that relationship for network design. The first half of this article traces the main ideas in graph theory over the past two centuries, which are at the basis of the mathematical approach to networks, and the second half is devoted to some very recent developments: computer network design and the connection between network dynamics and structure.

The Problem of the Königsberg Bridges

Network images can be quite striking. But one might question whether thinking about complex systems in terms of networks leads to more than pretty pictures. Ironically, the fundamentals of the theory of network structures were introduced by a blind mathematician.

It all began with the puzzle of seven bridges, an entertaining brainteaser for people who strolled through Königsberg, the Prussian city at the Baltic Sea, in the 18th century. The river Pregel divides the city into four land areas connected by seven bridges. The burghers of Königsberg wondered if one could visit all the four areas by crossing each bridge exactly once (see Figure 1).

The puzzle was solved in 1736 by Leonhard Euler, who at the time, was a mathematics professor in St. Petersburg. The power of Euler's solution lies not in the answer itself (which is negative) but in the way it was derived. Euler's revolutionary idea was to represent the pieces of land separated by bridges as the nodes (dots) *A*, *B*, *C*, and *D* and to represent the bridges as the edges (line segments) *a*, *b*, *c*, *d*, *e*, *f*, and *g*, connecting the nodes (see Figure 2). The structure formed by the set of nodes and edges, called a graph, is a simplified representation of the puzzle, encoding the relationships between the pieces of land and the paths of access between them (see Figure 2 inset). In this representation, the problem translates into the following one: Find a path that visits all nodes but passes through all edges exactly once. Obviously, the intermediate nodes must have an even number of incident edges (if one visits an intermediate node, one must also leave it).

Because the Königsberg puzzle has 4 (>2) nodes, all with an odd number of edges, there can be no such path.

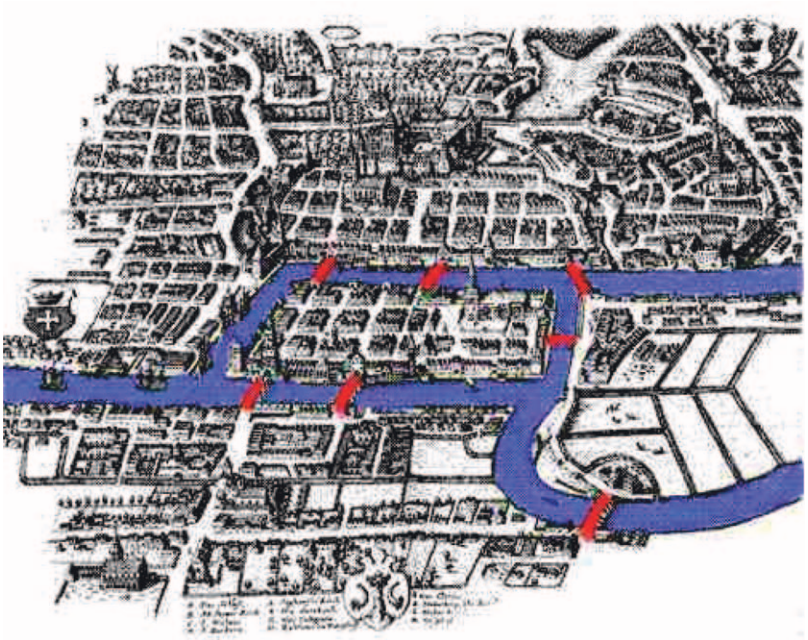


Figure 1. The Königsberg Puzzle
This woodcut shows the ancient Prussian city of Königsberg (now known as Kaliningrad) with its seven bridges across the river Pregel. The possibility of strolling across the city by crossing each bridge once only became the object of a famous brainteaser.

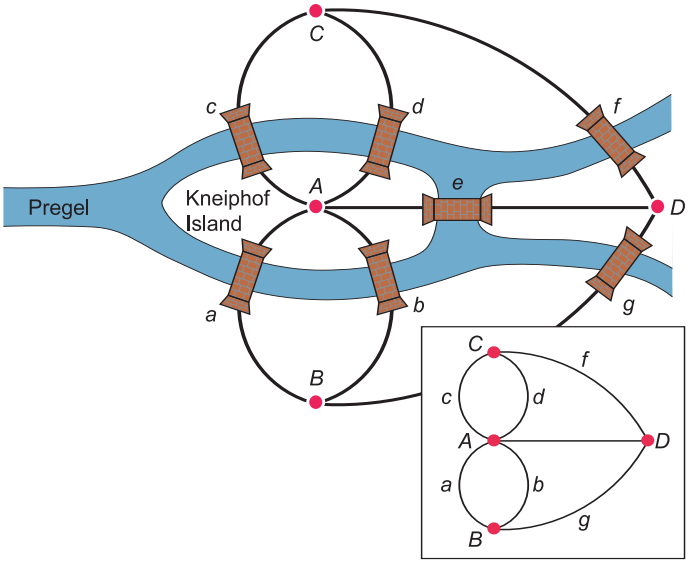


Figure 2. Euler's Solution to the Königsberg Puzzle
A simple representation of the problem by a graph helps realize that there is no such path that visits all nodes and passes through all seven edges exactly once.

Euler's representation of the relationships between a discrete set of entities as a graph led to the development of a particular type of mathematical nomenclature and ultimately to a new field of discrete mathematics called graph theory.

A Hard Problem: The Ramsey Numbers

For nearly 200 years, graph theory was concerned with topological and/or geometrical properties of small structures, or regular structures (such as a lattice). Then, the 1951 seminal paper by Ron Solomonoff and Anatol Rapoport (1951) and the 1959–1960 series of papers by Pál Erdős and Alfréd Rényi caused the rebirth of graph theory. These papers introduced the notion of a random graph and, more important, that of graph ensembles, which are sets of graphs that share a given property Q. To understand this notion, let us look briefly at the famous Party Problem and the Ramsey numbers. This problem, inspired from social interactions, is stated very simply:

What is the minimum number of guests, R, one should invite to a party that would surely have k people who all know each other or k who do not know each other (*at all*)?

For $k = 3$, it is easy to prove that $R(3) = 6$. We will use Euler's method: Let us denote the six people by the nodes A, B, C, D, E, and F. Let us represent the fact that two people know each other by drawing a red link (or edge) between them and use a blue edge to link two people who do not know each other. Since pairs of people either know each other or do not, the graph obtained is complete, which means that all possible edges are drawn—see Figure 3(a). Specifically, a complete graph with n nodes, denoted here by K_n, always has $n(n-1)/2$ edges. The graph theoretic version of the Party Problem is thus to determine the minimum number of nodes n, such that a complete graph with n nodes and with edges of two color always has at least one complete subgraph of k nodes with all edges of the same color. For $k = 3$, a complete subgraph is a monochromatic triangle.

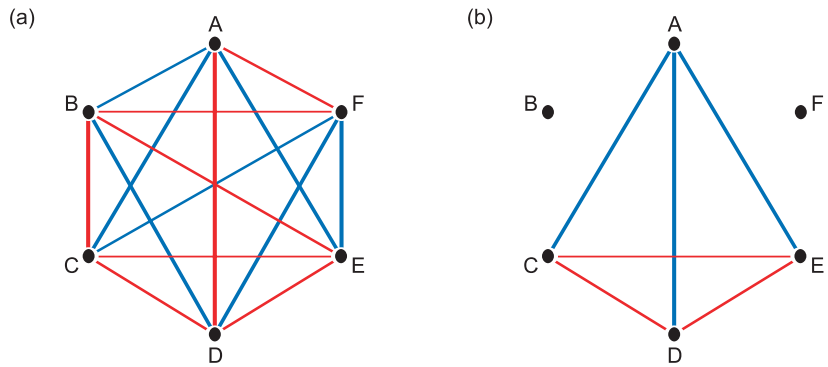


Figure 3. The Party Problem for $k = 3$
Six is the minimum number of people that always contains a group of three, all of whom either know each other (red links) or do not know each other (blue links). (a) A complete graph for six people is shown. Note that it contains the complete subgraph CDE. (b) This figure demonstrates that there is no way to draw a complete graph without constructing a complete single-color three-node subgraph within it. Suppose that blue links indicate that A does not know C, D, or E. In that case, CD, DE, and EC must be red links so that a complete blue subgraph should not be formed. But then, as shown in (b), those three red links form a complete subgraph, which means that C, D, and E know each other.

Leonhard Euler
Leonhard Euler, the most prolific mathematician of all times, was born in Switzerland in 1707 and spent his life in Berlin and St. Petersburg. *Opera Omnia* is an incomplete collection of his works that has 73 volumes, each over 600 pages in length.

If there are $n = 5$ people present, one can easily color a complete graph with no such triangle present, ruling out $n = 5$ by inspection.

For $n = 6$, however, there is always at least one such triangle. To prove this statement, let us assume the opposite, namely, that there can be no such triangles. Since for every node there are $n - 1 = 5$ incident edges but only two colors, there must be 3 edges of the same color incident on the node. For example, consider the edges AC, AD, and AE in Figure 3(b) to be the same color, for example, blue. Since the triangles ACE, ACD, and ADE cannot have all three of their edges of the same color, CE, CD, and DE must be red. Then CDE is a triangle all with the same color edges (red), a contradiction. Hence, $R(3) = 6$.

For $k = 4$, the answer is $R(4) = 18$, which is hard to prove. For $k = 5$ and higher, the answers are not known; only some bounds exist. Although we have no proof for $k = 5$, one might think that we would surely be able to use today's supercomputers to find the value of $R(5)$. However, as Bollobás,

an eminent graph theoretician, has stated (1998) "…a head-on attack by computers for $R(5)$ is doomed to failure . . . ." This failure is largely due to the combinatorial explosion in the number of ways we can draw a complete graph with n nodes using two colors for the edges: On the face of it, a computer would have to search a total of $2^{n(n-1)/2}$ such graphs for complete subgraphs with k nodes. For $k = 3$, when $n = R(3) = 6$, there are $2^{15} = 32{,}768$ complete graphs, for $k = 4$, the analytic solution gives $n = 18$, which means that there are 2^{153}, or approximately 1.46×10^{46} graphs. For $k = 5$, the best known bounds are $43 \leq R(5) \leq 49$, which would mean approximately 2^{903} to 2^{1176} graphs (or on the order of 10^{301} graphs). For $k = 5$ and $n = 43$, the "ultimate laptop" of Seth Lloyd (2000), which operates at the physical limit of computation (as determined by the speed of light, the Planck constant, and the gravitational constant), performing $f = 5.4258 \times 10^{50}$ operations per second, would have to work for at least 2.693×10^{213} years, a mighty long time (the age of the universe is estimated to be between 1.1×10^9 and 2×10^9 years).

So, can we hope ever to solve the Party Problem? The key idea is to understand how different colorings of K_n relate to one another via transformations, which would allow us to partition the set of two colorings of K_n into a smaller number of classes and, in the absence of a full mathematical theory, to program the computer to search for the monochromatic complete subgraphs on the set of classes instead of the full set. Although still unsolved, intense activity in this area led to a number of generalizations of this problem and to the development of a huge branch of mathematics, the Ramsey theory (Graham et al. 1990). That theory has a number of very deep results that go well beyond graph theory, affecting set theory in the form of partition calculus, combinatorics, ergodic theory, logic, analysis, algebra, geometry and computer science.

Pál Erdős (1913–1996)
Pál Erdős, who introduced the notion of random graphs, is probably the second most prolific mathematician of all times, having produced over 1500 publications with 507 coauthors.

Ultimately, the Party Problem suggests that, if we partition a set into a fixed number of classes, order must emerge for large enough sets. This principle is also illustrated by van der Waerden's theorem (Bollobás 1998), which states that, for a given k and p, if we partition the first w integers into k classes, we will always find a class that contains an arithmetic progression with p terms for large enough w. Problems like the Party Problem lead to a simple conclusion: In order to understand properties of graphs, one has to think in terms of ensembles of graphs that share a certain property, Q.

A Revolutionary Idea

The Hungarian mathematician Pál Erdős was one of the main pioneers of the ensemble approach. His complete disregard for the notion of possession and ownership and his habit of living out of a suitcase and visiting one mathematician friend after the next were symptoms, perhaps, of his total dedication to mathematics. Erdős is considered by many to be the second most productive mathematician of all times, after Euler. Possibly the greatest contribution of Erdős is his introduction of the probabilistic method in discrete mathematics. For graph theory, this means that, instead of asking for detailed properties of all graphs in an ensemble, we are asking for average properties, or the probability that a graph from an ensemble has the property Q. The probabilistic method was definitely not new when Erdős introduced it to discrete mathematics: By the end of the 19th century, Boltzmann, Gibbs, and others had laid down the foundations of equilibrium statistical mechanics, which is based on applying the probabilistic method to ensembles of microstates and characterizing macroscopic properties of the system by the properties of the "typical" microstates. This natural connection between statistical mechanics and graph theory is currently being exploited by some research groups worldwide, including the Statistical Physics of Infrastructure Networks team at Los Alamos. Besides the combinatorial explosion in the number of possible graphs (or states), there is a second strong reason that calls for the use of the probabilistic method: incomplete information. Real-world networks, as we will see from the following sections, are in many cases very dynamic, with new edges and nodes appearing and old ones disappearing as a result of stochastic processes. In addition, in some cases, it is hard, or even impossible, to identify precisely the graph structure at a given moment. Again, a good example is supplied by a problem related to social networks, namely, the Gossip Problem:

Suppose that person A in a set of N people has a very interesting piece of information or gossip. On average, how many acquaintances must every person in N have such that the gossip becomes known to all?

Since we do not know who knows whom, we determine the answer by considering all graphs of N people with the nodes representing the individuals and the edges representing the acquaintanceships, or social links, for transmitting gossip. In other words, if persons A and B are linked and one of them knows the gossip, we can assume the other knows it too. The answer to the Gossip Problem must be probabilistic in nature: It is the class of graphs having nodes with a certain average number of links and characterized by the property that everyone knows the gossip in the end. Erdős and Rényi came up with a rather surprising solution: *Once a node has on average one link, the gossip becomes known to all*! In the jargon of social scientists, the set of people represented by that graph forms a society. The class of graphs that Erdős and Rényi introduced and that helped give the answer is called random graphs, a subject with a huge mathematical literature. For a review, see the book by Bollobás (2001).

The Binomial Random Graph. Because we need it for later discussion, we introduce the binomial random graph *G(N,p)* and present some of its properties. $G(N,p)$ is a class of graphs with N vertices, whose edges are drawn at random and independently, according to a uniform distribution with probability p. Therefore, the average number of links incident on a node is $\lambda = p(N-1)$, or for large graphs, it is approximately $\lambda = pN$. Thus, according to the answer for the Gossip Problem, when $\lambda = 1$, or the probability for a node to have an edge is $p = p_c = 1/N$, a giant cluster, or giant component, emerges that contains most of the nodes, and the probability for a node not to belong to this cluster decreases exponentially fast for $p > p_c$. Physicists call this phenomenon percolation. Passing through p_c (by the process of increasing the average number of incident edges), the network suffers a drastic change, which is called a phase transition in the language of physics.

We now introduce one of the most important characteristics of random graphs, namely, their degree distribution. The degree of a node x is the number $k(x)$ of incident edges on that node. The degree distribution of the binomial random graph *G(N,p)* is the probability that the number of nodes X_k with degree k is y. In a $G(N,p)$, the probability of a node being connected to k specific other nodes and not connected to the rest of $N-1-k$ nodes is $p^k(1-p)^{N-1-k}$. Because the number of ways to connect those k nodes is equal to the binomial coefficient

$$\binom{N-1}{k},$$

the probability of a node having exactly k incident edges in *G(N,p)* becomes

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \quad . \qquad (1)$$

Note that, as edges are drawn incident to a node, that node will influence the number of edges around the other nodes, and thus, in principle, the distribution of X_k will not be exactly the same as if all the nodes were independent, and the calculation of the exact form of the degree distribution becomes a hard task. It was Bollobás (2001) who showed that, for large enough N, the nodes can be treated as if they were independent, and thus, with good approximation, the degree distribution of $G(N,p)$ is described by the binomial distribution in Equation (1). In the limit of $N \to \infty$ and $p \to 0$ such that $\lambda = pN =$ constant, the binomial goes into the Poisson distribution:

$$P(k) \approx e^{-\lambda} \frac{\lambda^k}{k!} \quad . \qquad (2)$$

Figure 4 shows a comparison between the formula in (2) and the measured degree distribution for a binomial graph of $N = 20{,}000$ nodes and a link probability $p = 20/N = 0.001$. It shows that, indeed, the approximation is good. The Poisson distribution *P(k)* has a "bell curve" shape, with a peak at $\lambda = pN$, and fast decaying tails. The degree of a node characterizes how a node "sees" its immediate neighborhood in the network. According to the formula in (2), if we keep $\lambda = pN$ a constant, while increasing the size of the network, the distribution of edges in the immediate neighborhood of a node becomes independent of N for large N. However, keeping $\lambda = pN$ a constant, means scaling the link probability p with $N-1$. The average node degree is

$$\langle k \rangle = \sum kP(k) = \lambda \ ,$$

and the standard deviation of the distribution around the average is

$$\sigma = \sqrt{\lambda} \quad .$$

This result shows that the binomial graph has a characteristic scale defined by λ.

Real-World Networks

The latest revolution in networks science happened toward the end of the 1990s, when powerful computers made it possible to gather and analyze data for systems containing a large

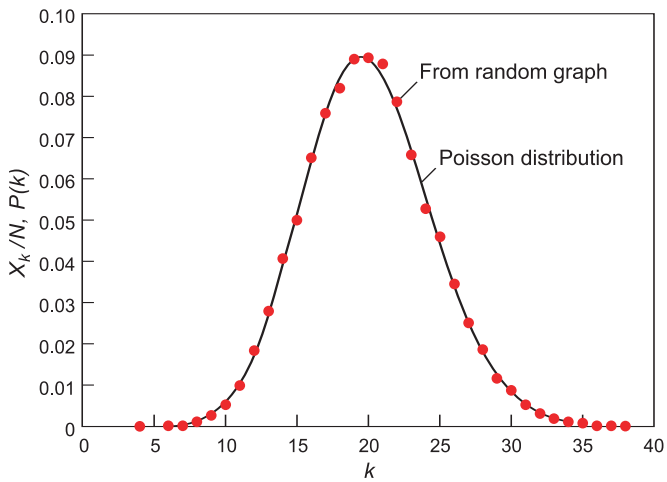


Figure 4. The Degree Distribution of the Binomial Random Graph
The red circles show the degree distribution, or X_k/N, the fraction of nodes with k links vs k, for a single instance of the binomial random graph $G(N,p)$ with the number of nodes $N = 2 \times 10^4$ and the probability for a link to exist between two nodes being $p = 10^{-3}$. The continuous black line is a plot of the Poisson distribution $P(k)$ in formula (2) with $\lambda = pN = 20$. Note the similarity between the two distributions.

number of components: from the World Wide Web and the Internet, phone call networks, networks of movie actors, large-company boards of directors, scientific collaboration networks, language networks, crime webs, epidemic networks, and the sex web to biological networks such as the metabolic network, protein interaction networks, cell-signaling, and food webs. The first important observation is that most of these networks are very different from the random graphs of Erdős and Rényi. In hindsight, this departure is not unexpected: In the random graphs of Erdős and Rényi, the edges are assumed to exist completely independently from each other, whereas in real-world networks, the existence of edges is typically conditioned by nonindependent processes, or constraints, such as spatial embedding and interaction range dependency. The real surprise is that, in spite of their diversity, real-world networks can be classified into a small number of different classes of graphs, each characterized by certain structural properties of the interaction topology in these systems. The most useful properties for this purpose are degree distributions, clustering, assortativity, and shortest paths.

Instead of listing the classes of these networks and enumerating their properties, we will discuss one ubiquitous class, the so-called scale-free networks, originally introduced by Albert-László Barabási. These networks have power-law degree distributions (see Figure 5), as opposed to the Gaussian or Poisson degree distributions of random graphs (for example, Figure 4). These real-world, scale-free networks include the network of movie actors, scientific collaboration networks, the sex web, the metabolic network in the cell (on all three levels of life—archaea, bacteria, and eukaryotes), the protein interactions network, the language network defined by synonyms (in which case, the nodes are the words, and the edges connect the synonyms), and virtually all large-scale information networks: the Internet (router and also autonomous domain level), the World Wide Web, some e-mail networks and phone call networks. Why do these real-world networks have similar degree distributions? Is there a universal mechanism that generates these structures? The first crucial observation is that, in most cases, these structures result from dynamic processes with a strong stochastic component, just like the random graph model of Erdős and Rényi. However, to deviate from the random graph model, the network evolution process must include stochastic dependency and bias. The question is then, "What stochastic processes will generate scale-free networks?"

Most current models generating scale-free networks identify a mechanism for network growth and evolution. Among the notable ones are the preferential attachment model of Albert and Barabási (2002), in which a newly arriving node connects to a node in the existing network with probability proportional to the current degree of the node in the network; the fitness-based network growth model of Caldarelli et al. (2002); the Chung-Lu model of power-law random graphs (2002); the model of the World Wide Web by Menczer (2002); the initial attractiveness model of Dorogovtsev et al. (2000); and others. Although these models produce graphs that have power-law degree distributions, they either have been built for a specific type of network (for example, the model by Menczer) or are mathematical abstractions in which the stochastic network-growth process has little to do with the actual, often quite complicated, evolution mechanism of the real-world network. The stochastic dynamics for the appearance and disappearance of Internet routers, which has many unknown factors, is another example. Most real-world networks are also strongly coupled to other networks or other large-scale complex systems, and thus, in order to identify the network evolution mechanism, one can-

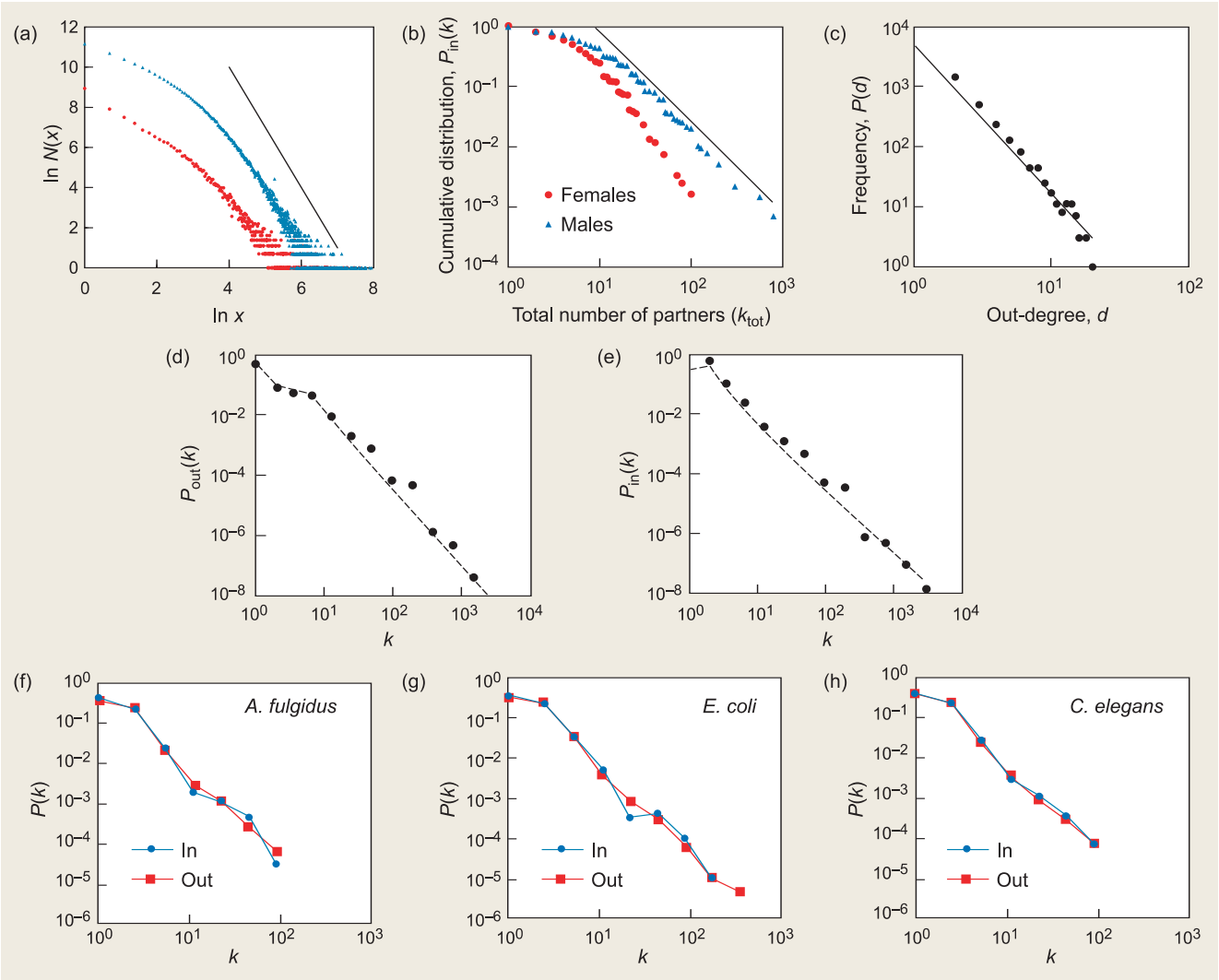


Figure 5. Degree Distribution of Various Scale-Free Networks

(a) Shown here is the cumulative degree distribution for citation networks (after Redner 1998); (b) the sex web (after Liljeros et al. 2001); (c) the Internet at the router level (after Faloutsos et al. 1999); (d) the in-link and (e) out-link degree distributions for the World Wide Web (after Albert et al. 1999); and (f) the metabolic networks for three species (after Jeong et al. 2000). [Plot (a) is courtesy of the *European Physical Journal B*. Plot (c) is courtesy of *Computer Communication Review*, ACM Publications 1999. Plots (b), (d), (e), (f), (g), and (h) are courtesy of *Nature*.]

not study these networks in isolation. To add to the complexity of the problem, the evolution of the network structure can depend on the dynamics or flow on the network. Most studies of complex networks have been static and structural as they try to identify their graph-theoretic properties. It has become clear, however, that, to solve even this problem, we must look at the full dynamics of the complex network, that is, at the flow on these structures and the coupling of the flow to the structural evolution.

The Problem of Epidemics. Before presenting some recent results that take into account the coupling between structure and dynamics, I will briefly mention an interesting and important real-world problem that is complex in the sense mentioned above. I am referring to epidemics, or disease propagation in living populations, a topic heavily studied at Los Alamos in the past decade. The usual, classic approach to epidemics imposes a number of assumptions that make analytic and numerical treatment relatively straightforward; however, at least in some cases, that approach may cause a departure from reality. One such assumption is uniform mixing, whereby the individuals of a population are assumed to come in contact with equal probability, independ-

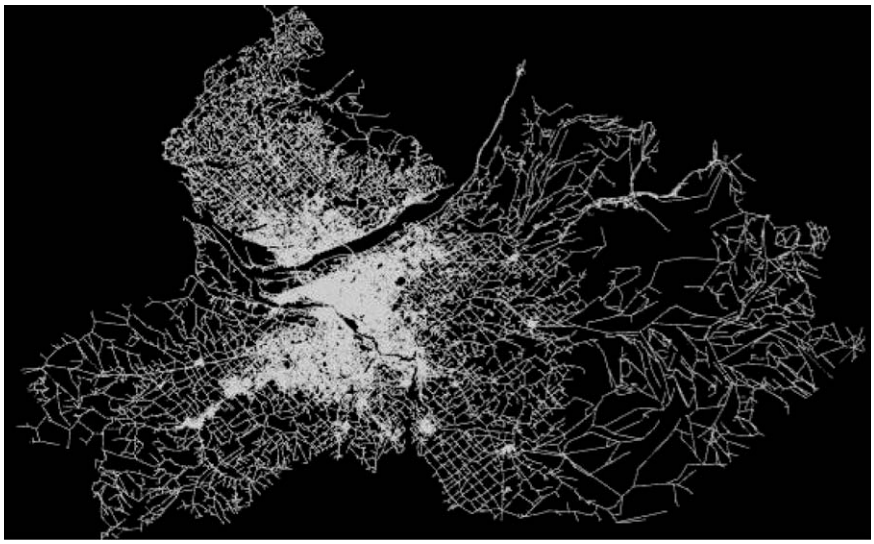


Figure 6. The Roadways of Portland, Oregon
The roadway network of Portland forms the substrate for a coupled complex dynamic network to simulate movements and disease transmission in this highly populated urban area. (This image is courtesy of the TRANSIMS Project at Los Alamos.)

ent of their locations. In order to relax this assumption, we observe that contact processes, such as disease transmission, are well localized in space and require that the two or more individuals be no farther apart than some typical distance characteristic of the disease transmission process. In heavily populated urban areas, disease is usually transmitted within such locations as buildings and mass transit areas (waiting areas and mass transit cars). Using census data and mobility diaries that specify the times of entrance and exit to and from a location for all locations that a specific person visited during the day, one obtains a graph that has the desired detailed resolution for contact patterns between people moving around in an urban area. This movement is largely constrained by the roadway network and the traffic on it. Since the roadway network is itself a complex network, the disease transmission problem is that of coupled complex dynamic networks (see Figure 6).

In this network, there are two types of nodes: people and locations, and an edge is drawn between a person and a location if that person visited that location during the day. The edge has a weight associated with it, called "timestamp," which is the union of time intervals during which the person was at that location. What can we learn, analyzing such a network, pertinent to disease outbreaks in a city? How can knowledge about this network be exploited to design effective vaccination and quarantine strategies? Conclusions for the specific case of Portland, Oregon, with approximately 1.6 million people and 181,000 locations can be found in Eubank et al. (2004).

Scale-Free Networks: Coincidence or Universality?

This section presents a different approach to understanding the emergence of the scale-free property for real-world networks. As mentioned previously, so far no one has found an obvious universal mechanism leading to power-law degree distributions for real-world networks. We actually suggest that, for a large class of networks (to be specified below), there is no universal evolutionary mechanism. Instead, the network structure evolves according to a selection principle that promotes the global efficiency of transport and flow processing through these structures (Toroczkai and Bassler 2004). In other words, regardless of the specific evolutionary mechanism, that mechanism works within the constraints of the selection principle. And the operation of the selection principle on evolution often results in scale-free networks.

Most real-world networks (except those that are defined by artificial associations) serve as transport substrates for various entities such as information, energy, material, and forces. Some networks have evolved spontaneously (without global design), and it makes sense to enquire whether their dynamics obey a selection principle toward some kind of optimal or efficient behavior. Such a principle would be analogous to the one of natural selection that shapes both the biological networks at the cellular level and the food web.

Looking at the Internet, we note that, if a router receives too much traffic and causes constant congestion of the packets, engineers will fix the problem locally by bringing up more routers or modifying the routing algorithm. Similarly, in social networks, if an acquaintance does not satisfy our expectations about some set of social norms, that link will "naturally" be dropped from our own social network.

To explore this trend toward efficiency more formally, we first need to define a flow process on the network. Among the most ubiquitous flow processes in Nature are those generated by local variations, or gradients, of scalar quantities. Particle concentration, temperature, electric or gravitational potential, and pressure are just a few examples. The gradient-induced flow processes include granular flow, fluid flow, electric current, diffusion

processes, heat flow, and so on. Naturally, the same local-gradient mechanism will generate flows in complex networks. Two less obvious local-gradient examples are diffusive load balancing schemes used in distributed computation (Rabani et al. 1998), which are also employed in packet routing on the Internet, and the reinforcement learning mechanism in social networks with competitive dynamics (Anghel et al. 2004). In the first example, a computer (or a router) will ask its neighbors on the network for their current job load (or packet load), and the router will balance its load with the neighbor that has the minimum number of jobs to run (or packets to route). In this case, the scalar is the negative of the number of jobs, or packets, at nodes, and the flow will be along the direction of the gradient of this scalar in the node's network neighborhood. In the second example, a number of agents/players in the social network play an interactive competition game with limited information. At every step of the game, each agent has to decide whose advice to follow before taking an action (such as placing a bet), in its circle of acquaintances (network neighborhood). Typically, an agent will try to follow the advice of that neighbor who in the past proved to be the most successful in predicting the game. That neighbor is recognized by using a scoring mechanism, which is the simplest form of reinforcement learning: Every agent has a success score that changes in time, coupled to the game's evolution. An agent will follow the advice of that agent who has the highest score in its network neighborhood at that moment (Anghel et al. 2004). In this case, the scalar is the past success score of the agents, and an agent will act based on the information received along the link that is in the direction of the gradient of this scalar.

To construct a simple and general model of a transport process, we assume that there are N nodes and that the transport takes place on a fixed substrate network $S(V, E)$, where V is the node set and E is the edge set that describes the interconnections of the nodes. Associated with each node i, there is a scalar h_i that describes the "potential" of the node. Then a gradient network G can be constructed as the collection of directed links that point from each node to the nearest neighbor on the substrate network S that has the highest potential (see Figure 7). Thus, only one directed link points away from each node in G, and G consists of N directed links. Note that, if the potential of a node is higher than the potential of all its nearest neighbor nodes, the gradient link of that node is a loop that points back to itself ("self-loop"). In general, the potential for each point can evolve in time, and as a result, the gradient network G will be time dependent. If we furthermore assume that all links have the same conductance, or transport properties, the gradient network will describe the instantaneous substructure carrying the maximum flow. Consequently, we can hope to use gradient networks as a tool to analyze the flow efficiency or susceptibility to jamming on the corresponding substrate networks.

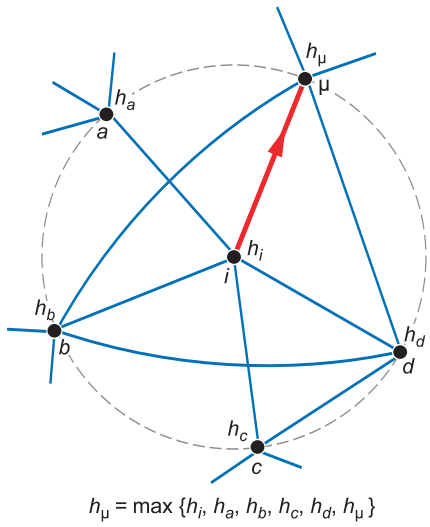


Figure 7. Definition of a Gradient Edge

The gradient edge is a directed link from node *i* to that neighbor on the substrate graph that has the largest value of the scalar in the neighborhood of *i*. If *i* has the largest value, then the gradient edge is a self-loop.

Note that, if there are two or more nodes in the network neighborhood of a node i that share the maximum value, the gradient in i is called degenerate. If each neighborhood has only one maximum, it is called nondegenerate, and is easier to analyze. In the discussions below, we will restrict ourselves to the nondegenerate condition, which is easily realized if, for example, the scalars are continuous random variables. Since every node has exactly one gradient direction from it (even f it is a self-loop), G has exactly N nodes and N edges (and there is at least one self-loop, corresponding to $\max_i\{h_i\}$). A simple but very important property of nondegenerate gradient networks is that they form forests, that is, each gradient network is a collection of tree graphs containing no loops (except for self-loops). We can therefore hope to analyze network flow processes using the techniques of statistical mechanics that have been well developed for treelike structures.

Gradient Networks on Random vs Scale-free Networks. Let us first consider a gradient network for a random graph substrate S. In particular, we choose for S the binomial random graph, $G(N,p)$ consisting of N nodes, each pair of nodes being linked with probability p. We next assume that the scalar potentials of the various nodes are independent random variables identically distributed according to a distribution $\eta(h)$. The distribution of

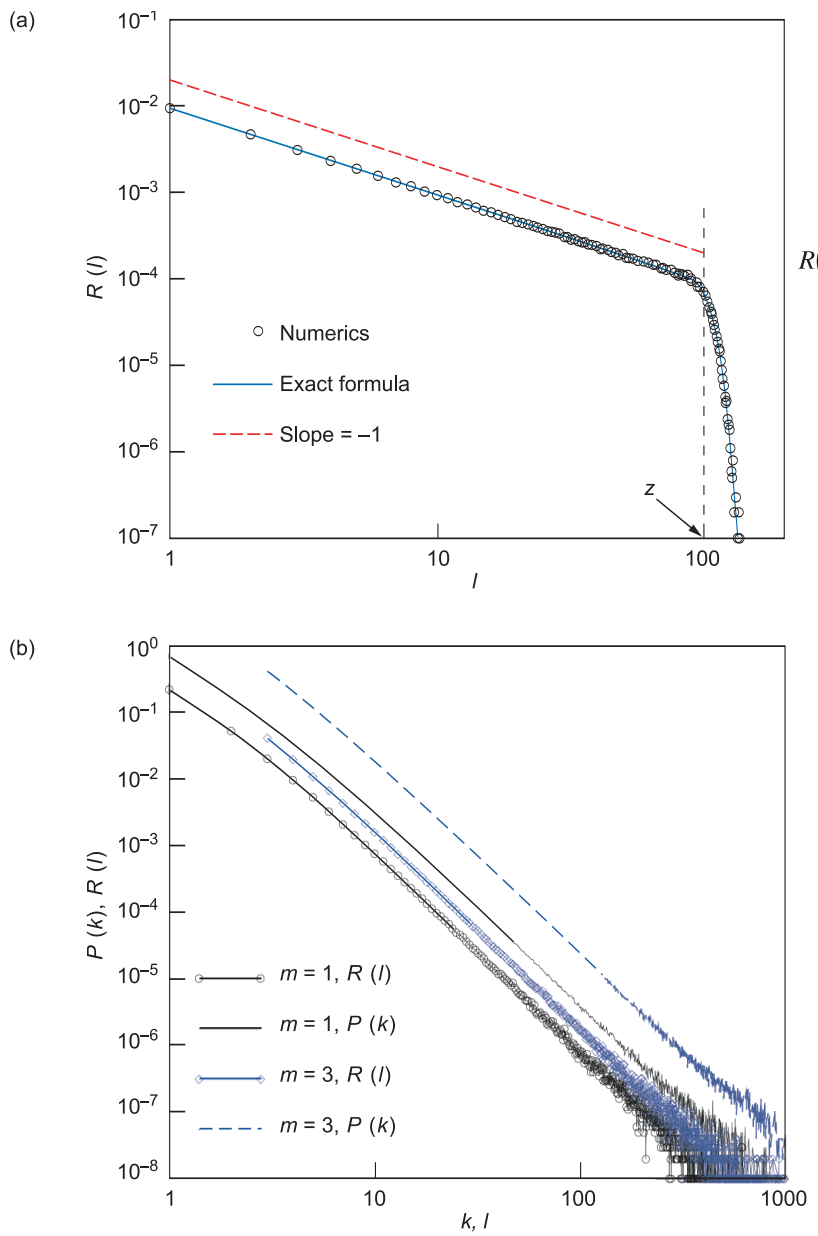


Figure 8. Gradient-Graph Degree Distributions for Random and Scale-Free Substrate Networks

(a) The in-degree distribution is shown for the substrate binomial random graph $G(N,p)$, where $N = 1000$, and $p = 0.1$ ($z = 100$). The numerical values are obtained after averaging over 10^4 sample runs. (b) The in-degree and degree $P(k)$ distributions are for the substrate Barabási-Albert scale-free graph with parameter m ($m = 1, 3$). In this case $N = 10^5$, and the average is performed over 10^3 samples.

the number of links l pointing to each node, the so-called in-degree distribution $R(l)$ of the gradient network G, can be exactly calculated, and it yields the following expression:

$$R(l) = \frac{1}{N}\sum_{n=0}^{N-1}\binom{N-1-n}{l} \times\left[1-p(1-p)^n\right]^{N-1-n-l}\left[p(1-p)^n\right]^l . \quad (3)$$

Thus, this in-degree distribution is independent of the particular form of the distribution for the scalar potentials $\eta(h)$. It is possible to show that in the limit $N \to \infty$ and $p \to 0$ such that $Np = \lambda =$ constant $>> 1$, the expression in Equation (3) becomes the power law $R(l) \approx 1/(\lambda l)$, with a finite-size cutoff at $l_c = z$; refer to Figure 8(a). Therefore, in this limit, gradient networks are scale-free graphs (up to their cutoff)! This power-law degree distribution for the gradient network is a rather surprising result because, in the same limit, the substrate graph S is a binomial random graph having a Poisson degree distribution with a well-defined average degree λ (setting the scale of the substrate graph), as well as rapidly decaying tails.[1]

If, instead, the substrate network S is a scale-free graph, the gradient graph will still have a power-law degree distribution. Figure 8(b) compares degree distributions $P(k)$ for

[1]A similar finding was reported by Lakhina et al. (2003), who repeated on binomial random graphs the trace-route measurements used to sample the structure of the Internet. Lakhina and colleagues found that the spanning trees obtained in this way have a degree distribution that obeys the $1/k$ law. Later, Clauset and Moore (2003) have presented an analytical approach to derive the $1/k$ law. This approach suggests the possibility of mapping between graphs generated by trace-route sampling and gradient networks. Although it is not an exact mapping, a close connection can indeed be made by interpreting trace-route trees as suitably constructed gradient networks.

scale-free substrate networks, which we generated by the Barabási-Albert network with parameter m (minimum degree, see Albert and Barabási 2002) with the in-degree distributions for the corresponding gradient networks. One immediate conclusion is that the gradient network is the same type of structure as the substrate. In this case, it is a scale-free (power-law) graph with the same exponent.

Flow Properties on Random vs Scale-Free Networks. Using the properties of gradient networks, we can define a transport characteristic related to congestion or jamming in the substrate network. In particular, we compare the average number of nodes with in-links with the average number of nodes with out-links. If $N_l^{(\mathrm{in})}$ denotes the number of nodes with l in-links, the total number of nodes receiving gradient flow will be

$$N_{\mathrm{receive}} = \sum_{l \geq 1} N_l^{(\mathrm{in})} \quad .$$

The total number of gradient out-links is simply $N_{\mathrm{send}} = N$ because every node has exactly one out-link. Naturally, the ratio $N_{\mathrm{receive}}/N_{\mathrm{send}}$ will be related to the instantaneous global congestion in the network. The smaller the number of nodes receiving the flow (given the same number of senders), the more congestion is in the substrate network at that instant. If the flow received by a node requires a nonzero processing time (such as routing of a packet by the router), a small ratio of $N_{\mathrm{receive}}/N_{\mathrm{send}}$ translates into large delay times and thus inefficient flow processing. Let us define the congestion (or jamming) factor as follows:

$$J = 1 - \left\langle \left\langle \frac{N_{\mathrm{receive}}}{N_{\mathrm{send}}} \right\rangle_n \right\rangle_h = R(0) \ , \qquad (4)$$

where $\langle\ \rangle_n$ means averaging over the disorder in the network structure and $\langle\ \rangle_h$ means averaging over the randomness in the scalar field. The value of J is always between 0 and 1, with $J = 1$ corresponding to maximal congestion and $J = 0$ corresponding to no congestion. Note that J is a congestion pressure characteristic generated by gradients rather than an actual throughput characteristic. For a binomial random substrate network $G(N,p)$, we use Equations (3) and (4) to obtain the corresponding jamming factor:

$$J^R(N,p) = \frac{1}{N} \sum_{n=1}^{N-1} \left[1 - p(1-p)^n \right]^{N-1-n} \quad . \qquad (5)$$

In the scaling limit $N \to \infty$ and $p =$ constant, the jamming factor assumes the asymptotic behavior

$$J^R(N,p) = 1 - \frac{\ln N}{N \ln\left(\frac{1}{1-p}\right)} \times \left[1 + O\left(\frac{1}{N}\right) \right] \to 1 \quad .$$

That is, the random graph becomes maximally congested. It is easy to show that, in the other limit, when $z = Np >> 1$ is kept constant while $N \to \infty$,

$$J^R(z) \approx 1 - \frac{\ln z}{z} - \ldots \to 1 \quad .$$

Once again, the random graph asymptotically becomes maximally congested, or jammed.

For scale-free networks, however, the conclusion about jamming is entirely different. We find that the jamming coefficient J becomes independent of N, and it is always a constant less than unity for large networks. In other words, scale-free networks are not prone to maximal congestion. (This is true for all power-law networks for which the average degree does not grow with N.) Figure 9 shows the congestion factors as a function of network size for random and scale-free substrate networks. Many real-world networks evolve more or less spontaneously (for example, the Internet or the World Wide Web), and they can reach sizes of about 10^8 nodes. At such large N, the scaling limit studied above applies, and random networks have maximal congestion. Thus, such substrates are very inefficient for flow processing. Scale-free networks, on the other hand, have congestion that stays bounded away from unity as the number of nodes grows very large, and they are therefore much more efficient substrates for transport and flow processing. Thus, it appears that the scale-free property of many real-world networks is not accidental. Topology may develop quite naturally from a selection rule that tends to maximize the global efficiency of the flow along the network.

Small-World Magic: Synchronized Computing Networks

We have seen that many real-world interactions are mediated across complex network topologies and that the structure and dynamics of those complex networks are becoming better understood. It is therefore natural to wonder whether network concepts can be put to practical use. For example, can those concepts help us design systems that exhibit certain desired properties? In this last section of the article, I will show how complex network concepts were used to solve a problem in distributed, or parallel, computation.

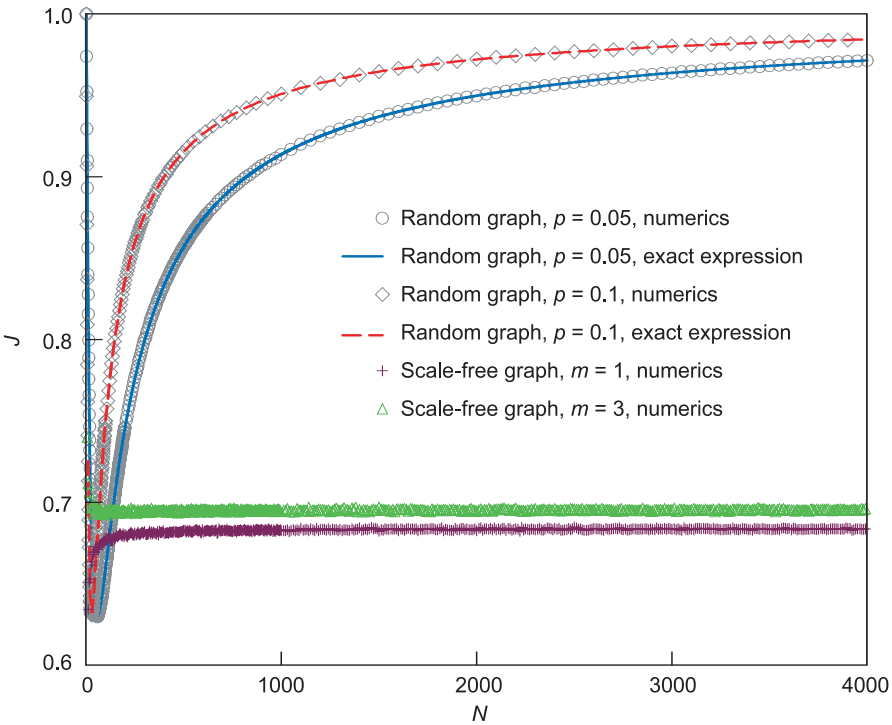


Figure 9. Congestion Factors for Random and Scale-Free Substrate Networks

Congestion factors are shown as a function of size for random graphs and scale-free networks. For random binomial graph substrates, the jamming coefficient tends to unity with increasing network size, indicating that these networks will become extremely congested in this limit. For scale-free substrates, however, the congestion factor becomes independent on the network size, and thus arbitrarily large networks can be considered without increasing their congestion level.

We consider the class of systems made of a large number of interacting elements or individuals, each having a finite number of attributes, or local state variables, that can assume a countable number (typically finite) of values. The dynamics of the local state variables are discrete events occurring in continuous time, and the interactions between individuals, or elements, have a finite range. There are many examples of such systems: magnetic systems, epidemics, some financial markets, wireless communications, queuing systems, and so on. Virtually all agent-based systems can be considered to belong to this class of discrete-event complex systems. Often, the dynamics of such systems is inherently stochastic and asynchronous. Simulating the systems is nontrivial, and in most cases, the complexity of the problem requires the use of distributed computer architectures. These problems define the field of parallel discrete-event simulations (PDES).

Conceptually, the computational task is divided among N processing elements (PEs), each of which evolves the dynamics of the allocated piece of the system. Because of the interactions among the individual elements of the real system (spins, atoms, packets, calls, and so on), the PEs must coordinate with a subset of other PEs during the simulation.

At present, large parallel computers for performing PDES have thousands of nodes and soon will have tens of thousands: the Nippon Electric Company's 5120-node Earth simulator producing 35.86 teraflops, the 8192-node Q-machine at Los Alamos with 13.88 teraflops, Virginia Tech's X machine, which is a 2200-node apple G5 cluster with 10.28 teraflops, and so on. IBM is currently building the Blue Gene/L parallel computer with 360 teraflops and 65,000 nodes. Blue Gene/P, the next-generation computer, is expected to surpass the petaflop barrier in 2006.

The design of efficient, scalable update schemes for performing PDES on these large parallel computers is a rather challenging problem because the simulation scheme itself becomes a complex system whose properties are hard to deduce using classical methods of algorithm analysis. Korniss et al. (2003) introduced a less conventional approach to analyzing the efficiency and scalability of parallel discrete-event simulation schemes. The authors constructed an exact mapping between the parallel computational process itself and a nonequilibrium surface growth model. As a result, questions about efficiency and scalability can be mapped into certain topological properties of this nonequilibrium surface. Then, using methods from statistical mechanics, we can solve the scalability problem of the computation PDES schemes. We now briefly sketch the scalability problem and its solution.

In order to simulate the dynamics of the underlying system, the PDES scheme must track the physical-time variable of the complex system. In case of asynchronous dynamics on distributed architectures, each PE generates its own physical (also called virtual) time τ, which is the physical time variable of the particular computational domain handled by that PE. Because of the varying complexity of the computation at different PEs, at a given wall-clock instant, the simulated, or virtual, times of the PEs can

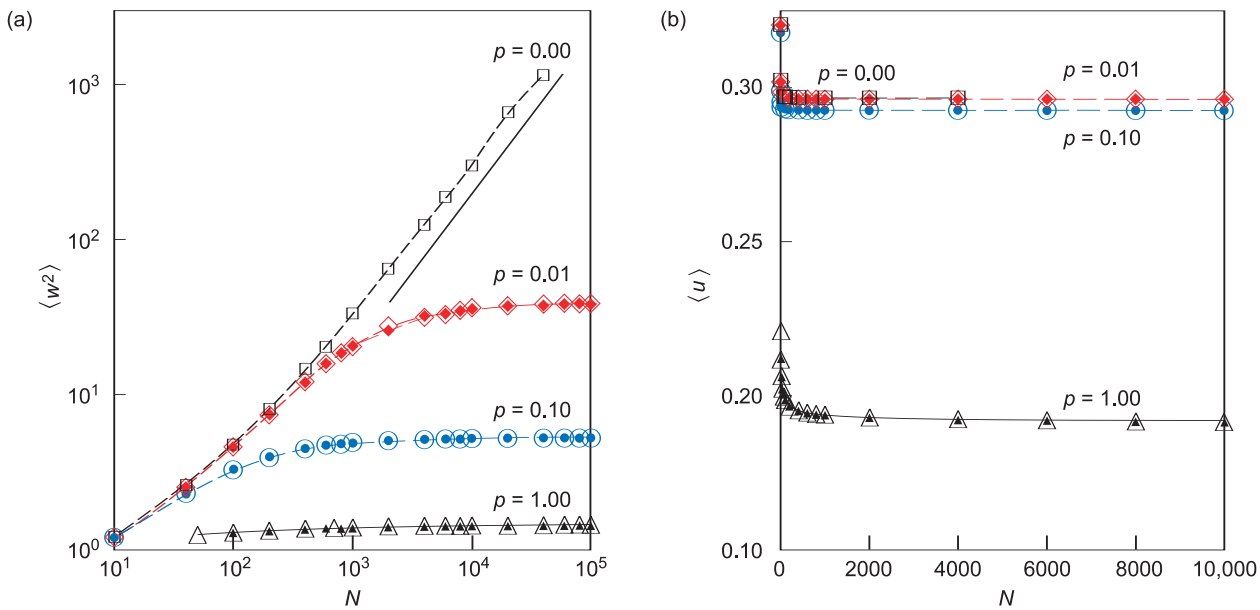


Figure 10. A Fully Scalable Small-World PDES Scheme

The small-world PDES scheme is fully scalable. By introducing more shortcuts into the communication network (increasing *p*), the algorithm becomes measurement scalable (a), and it stays computationally scalable (b).

differ, a phenomenon called "time horizon roughening." Let us denote the virtual time at PE_i measured at wall-clock time t by $\tau_i(t)$. The set of virtual times $\{\tau_i(t)\}_{i=1}^N$ forms the virtual time horizon of the PDES scheme after t parallel updates. In conservative PDES schemes, a PE will perform its next update only if it can obtain the correct information from its neighbors to evolve the local configuration (local state) of the underlying physical system it simulates without violating causality. Otherwise, it idles. Specifically, the PE_i can only update (become "active") at wall-clock instant t if

$$\tau_i(t) \le \min_{j \in <i>} \left\{\tau_j(t)\right\} . \qquad (6)$$

That is, the PE's virtual time is a local minimum among the virtual times of its neighboring PEs (specified as the set <*i*>). Once the PE at site i can update, it will advance its local simulated time to the new value $\tau_i(t + 1)$, and the process is repeated for all active sites, generating the dynamics of the virtual time horizon $\{\tau_i(t)\}_{i=1}^N$. The average of the time horizon after t parallel steps is obviously

$$\bar{\tau}(t) = \frac{1}{N}\sum_{j=1}^{N} \tau_j(t)$$

Thus, the rate of progress of the time horizon average, or the average utilization of the PEs $\langle u(t)\rangle = \langle\bar{\tau}(t + 1)\rangle - \langle\bar{\tau}(t)\rangle$ is proportional to the number of nonidling, or active, PEs. The average $\langle\cdot\rangle$ is taken over the stochastic event dynamics. The PDES scheme is computationally scalable if there is a constant $c > 0$, such that

$$\langle u(\infty)\rangle = \lim_{t,N\to\infty} \langle u(t)\rangle > c \quad . \qquad (7)$$

That is, the average rate of progress of the time horizon does not vanish even after very long times, as the simulated system size and, therefore, the number of PEs are taken to infinity.

We solved this computational scalability problem by drawing an analogy with the statistical mechanics of nonequilibrium surface-growth processes. Thin films are grown on solid substrates by deposition of atoms or molecules from surrounding vapors. Because the vapors are fairly hot, the atoms reaching the solid surface follow a stochastic path until they are incorporated into the surface, typically in an irregular fashion. The resulting thin film has mounds and valleys that can be described by the fluctuations of the local height variable $h(x,t)$ of the film measured from the surface of the substrate. Using an approach based on the Langevin equation, physicists have developed extended theoretical machinery to describe the statistics of the fluctuations of the variable $h(x,t)$. The simulated time variable $\tau_i(t)$ in the computational scalability problem

behaves much like the surface height variable $h(x,t)$ in that $\tau_i(t)$ evolves according to the stochastic update dynamics of the PDES scheme with the index i of the PE corresponding to x in the height variable. In many large complex systems, the dynamics of the stochastic events can be characterized by a Poisson distributed stream. This means that, when simulating such systems, the updates at individual PEs correspond to adding height increments that follow a Poisson distribution. Using statistical mechanics methods to analyze the resulting surface-growth model, one can show that the fluctuations of the virtual time horizon in the continuum limit can be described by the so-called Kardar-Parisi-Zhang (KPZ) equation of surface growth:

$$\frac{\partial \hat{\tau}}{\partial t} = \frac{\partial^2 \hat{\tau}}{\partial x^2} - \lambda\left(\frac{\partial \hat{\tau}}{\partial x}\right)^2 + \eta(x,t) \ , \qquad (8)$$

where $\hat{\tau}$ is a coarse-grained form of the virtual-time variable and η is a white noise term.[2] We then use the KPZ equation to verify that the utilization of the PEs satisfies Equation (7). The existence of a constant $c > 0$, as in (7), is the result of the slope-slope correlations of the surface being short ranged and not scaling with N. Our numerical evaluation of this constant yields $c = 0.2461 \pm (7 \times 10^{-6})$, which shows that the basic conservative PDES scheme is indeed computationally scalable.

There is, however, a fundamental problem with the basic PDES update scheme. The KPZ equation for the time horizon fluctuations predicts that the average spread of those fluctuations, $w^2(N,t)$, diverges with an increasing number of processing elements N in the long time limit ($t \to \infty$). Therefore, if we try to measure a global property of simulated system at a given simulated time τ and wait until all processors have simulated their local state corresponding to that time, the waiting period in wall-clock time would diverge with the number of processing elements! In other words, even though a parallel computer with infinitely many processing elements can simulate the dynamics of an infinitely large system at nonzero speed (computational scalability), the basic PDES scheme could not produce a single measurement of the global state of the system! The basic conservative scheme is computationally scalable but measurement nonscalable.

How can we surmount this problem? Can the PDES scheme be modified such that the new update scheme is also measurement scalable? The answer is affirmative, and the key to the solution is the notion of the small-world property of complex networks (Korniss et al 2003).

In order to decorrelate the fluctuations in the time horizon, we modify the update topology in the following way: for every node i, we assign a randomly chosen communication link, $r(i)$. According to its definition, the resulting communication topology (a regular lattice plus random links) forms a small-world network. When a node is allowed to update—its virtual time satisfies the condition in (6)—it will make, with probability p, an extra check for the condition $\tau_i(t) \le \tau_{r(i)}(t)$ and update if that condition is satisfied. With probability $1 - p$, it will make this extra check and thus behave as the basic PDES scheme. Here p has the role of a tuning parameter: For p = 0, we have the basic PDES scheme, whereas p = 1 corresponds to the fully scalable small-world PDES scheme. Note that these extra checks do not affect the correctness of the simulation, and causality is preserved in just the same way. These checks only synchronize the PEs. Using the same coarse-graining methods as for the basic PDES scheme, we now find that the time horizon fluctuations are described by

$$\frac{\partial \hat{\tau}}{\partial t} = -\gamma(p)\hat{\tau} + \frac{\partial^2 \hat{\tau}}{\partial x^2} - \lambda\left(\frac{\partial \hat{\tau}}{\partial x}\right)^2 + \eta(x,t) \ , \qquad (9)$$

with $\gamma(0) = 0$ and $\gamma(p) > 0$ for $0 < p \le 1$. This equation differs from Equation (8) in the strong damping term, $-\gamma(p)\hat{\tau}$, which is ultimately responsible for the nondivergence of the average spread, and thus the new update scheme is measurement scalable as shown in Figure 10.

Concluding Remarks

The list of problems, challenges, and applications that I presented above is rather biased toward my particular research interests, and it is not, by far, exhaustive of this area. My goal was to give the reader a feeling for the type of complexity one encounters when dealing with networks. I also wanted to show that this is a novel area with many interesting and potentially powerful applications awaiting discovery. ■

Further Reading

Albert, R., and A.-L. Barabási. 2002. Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.* **74**: 47.

Albert, R., H. Jeong, and A.-L. Barabási. 1999. Diameter of the World-Wide Web. *Nature* **401**: 130.

Anghel, M., Z. Toroczkai, K. E. Bassler, and G. Korniss. 2004. Competition-Driven Network Dynamics: Emergence of a Scale-Free Leadership Structure and Collective Efficiency. *Phys. Rev. Lett.* **92**: 058701.

[2] The deterministic part of this equation is easily related to the well-known Burgers equation through the slope variables $\hat{\phi} = \partial\hat{\tau}/\partial x$.

Bollobás, B. 1998. *Modern Graph Theory*: *Graduate Texts in Mathematics*. Vol. 184. Edited by F. W. Gehring, and S. Axler. New York: Springer-Verlag.

———. 2001. *Random Graphs*. Second Edition. Cambridge; New York: Cambridge University Press.

Caldarelli, G., A. Capocci, P. De Los Rios, and M. A. Muñoz. 2002. Scale-Free Networks from Varying Vertex Intrinsic Fitness. *Phys. Rev. Lett.* **89**: 258702.

Chung, F., and L. Lu. 2002. Connected Components in Random Graphs with Given Expected Degree Sequences. *Ann. Comb.* **6**: 125.

Clauset, A. and C. Moore. 2004. Traceroute Sampling Makes Random Graphs Appear to Have Power Law Degree Distributions. [Online]: http://arxiv.org/abscondmat/0312674

Dorogovtsev, S. N., J. F. F. Mendes, and A. N. Samukhin. 2000. Structure of Growing Networks with Preferential Linking. *Phys. Rev. Lett.* **85** (21): 4633.

Eubank, S., H. Guclu, V. S. A. Kumar, M. V. Maratho, A. Srinivasan, Z. Toroczkai, and N. Wang. 2004. Modelling Disease Outbreaks in Realistic Urban Social Networks. *Nature* **429**: 180.

Faloutsos, M., P. Faloutsos, and C. Faloutsos. 1999. On Power-Law Relationships of the Internet Topology. *Comput. Commun. Rev.* **29** (4): 251.

Graham, R. L., B. L. Rothschild, and J. H. Spencer. 1990. *Ramsey Theory*. Second Edition. New York: John Wiley and Sons, Inc.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. 2000. The Large-Scale Organization of Metabolic Networks. *Nature* **407**: 651.

Kilian, J., Y. Rabani, A. Sinclair, and R. Wanka. 1998. Local Divergence of Markov Chains and the Analysis of Iterative Load Balancing Schemes. In *Proceedings of the 39th Annual Symposium on the Foundations of Computer Science (FOCS)*. p. 694. Los Alamitos, CA: IEEE Computer Society.

Korniss, G., M. A. Novotny, H. Guclu, Z. Toroczkai, and P. A. Rikvold. 2003. Suppressing Roughness of Virtual Times in Parallel Discrete-Event Simulations. *Science* **299** (5607): 677.

Lakhina, A., J. W. Byers, M. Crovella, and P. Xie. 2003. Sampling Biases in IP Topology Measurements. In *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies* (*INFOCOM*). p. 332. New York: IEEE.

Liljeros, F., C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg. 2001. The Web of Human Sexual Contacts. *Nature* **411**: 907.

Lloyd, S. 2000. Ultimate Physical Limits to Computation. *Nature* **406**: 1047.

Menczer, F. 2002. Growing and Navigating the Small World Web by Local Content. *Proc. Natl. Acad. Sci. U.S.A.* **99** (22): 14014.

Redner, S. 1998. How Popular Is Your Paper? An Empirical Study of the Citation Distribution. *Eur. Phys. J. B*, **4** (2): 132.

Solomonoff, R., and A. Rapoport. 1951. Connectivity of Random Nets. *Bull. Math. Biophys.* **13**: 107.

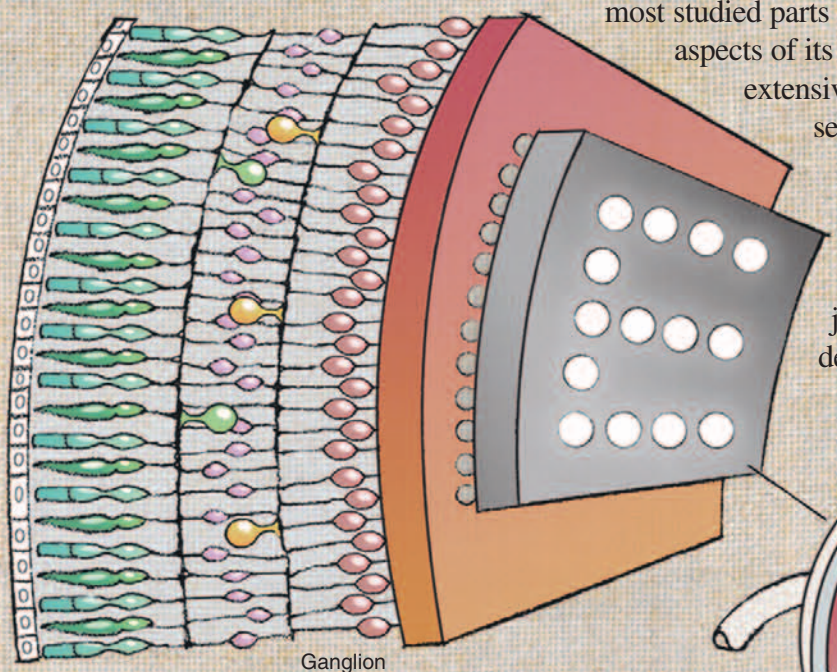
Toroczkai, Z., and K. E. Bassler. 2004. Jamming is Limited in Scale-Free Systems. *Nature* **428**: 716.

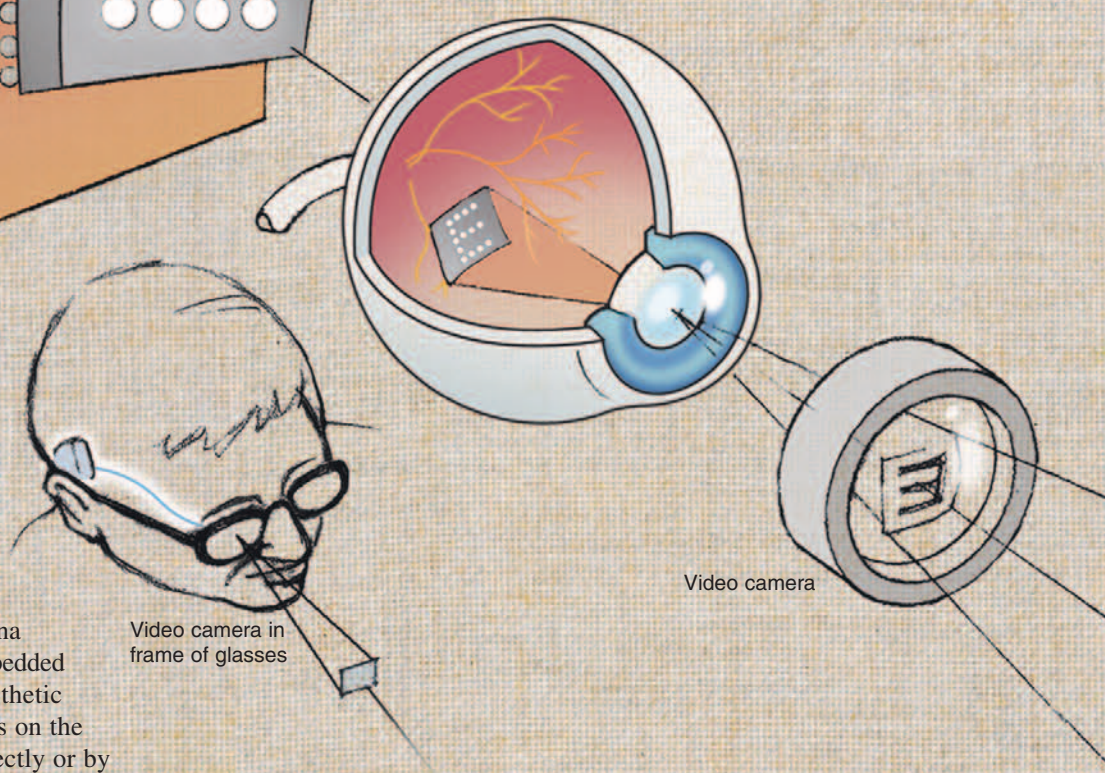
For further information, contact Zoltán Toroczkai (505) 667-3218 (toro@lanl.gov).

Models of the Retina with Application to the Design of a Visual Prosthesis

Garrett T. Kenyon, John George, Bryan Travis, and Krastan Blagoev

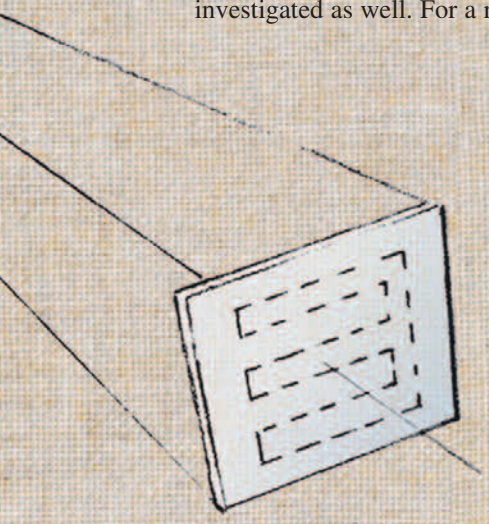
The retina, the neuronal layer at the back of the eyeball, is one of the most studied parts of the central nervous system, yet many aspects of its function remain a mystery. Does the retina extensively process incoming light signals before sending them to the brain, or does it serve primarily as a passive conduit? Experiments measuring retinal responses and subsequent detailed computer modeling have led to a novel conjecture, which may be significant in designing a sophisticated retinal prosthesis.





Clinical studies are under way to test the first primitive visual prosthetics. Prototype devices consist of microelectrodes implanted in the retina and stimulated by a video system embedded in a special pair of glasses. The prosthetic stimulates the array of ganglion cells on the inner surface of the retina either directly or by activating their synaptic inputs, thereby causing them to fire action potentials that propagate along the optic nerve to processing centers in the brain. Creating firing patterns that match those produced in the healthy retina by natural visual stimuli is the foremost challenge confronting the development of a retinal prosthesis.

Some forms of adult-onset blindness are characterized by a massive loss of photoreceptors but a relative sparing of fibers in the optic nerve. Recent clinical studies suggest that patients suffering from such visual impairments could benefit from a prosthetic device capable of stimulating the remaining retinal neurons and thereby mimicking the function of the missing rods and cones. A retinal prosthesis is illustrated conceptually on the opening page of this article. The light transduction role of the damaged or missing photoreceptors is performed by a video camera attached to a pair of specially configured eyeglasses worn by the patient. The video image is processed and then transmitted, through a cable or some form of wireless telemetry, to a multielectrode array attached to the inside surface of the retina. Stimulating the multielectrode array in an approximately one-to-one spatial correspondence with the video image will hopefully produce patterns of neural activity in the optic nerve similar to those produced by an undamaged retina during normal vision. Preliminary studies, in which a crude prototype of the above design was used, yielded encouraging results (Humayun et al. 2003). Other strategies for a retinal prosthesis—particularly the insertion of a standalone photodiode array between the retina and the back of the eye—are being investigated as well. For a recent review, see Margalit et al. (2002).



A consortium of DOE laboratories is working on a number of difficult technical problems that must be solved before a functional retinal prosthesis becomes widely available. Here, we describe computer modeling studies that have two goals regarding the optimal design of a retinal prosthesis: (1) to discover how visual information is processed and encoded by retinal circuitry, discussed in the main article, and (2) to improve our understanding of how retinal components, at the level of individual cells and across interconnected circuits, are activated by specific spatiotemporal patterns of electrical stimulation, discussed in the box, "Modeling Stimulation by a Retinal Prosthesis" on page 122. Understanding how the retina encodes visual information and how surviving elements in the diseased retina react to external stimulation is critical to achieving maximal therapeutic benefit from a prosthetic device.

Attempts to develop computer models of the retina benefit greatly from a large existing knowledge base. The anatomy and physiology of the retina have been extensively studied, especially in comparison with many other parts of the central nervous system. Moreover, the inputs and outputs of the retina have been well characterized. Because it receives no major feedback from the brain, the retina can be studied as a standalone circuit. Thus, we use experimental data to constrain our computer simulations of the retina to an extent not possible when modeling more central brain areas. Nevertheless, our studies may provide valuable insights into the design of neural prosthetics for other, less-accessible brain regions and may suggest new image-processing strategies for computer-vision systems.

The Retina

The retina consists of several layers of cells at the back of the eye that collectively are responsible for the transduction and preprocessing of visual signals (Figure 1). In photomicrographs of retinal cross-sections, several processing layers can be distinguished. In patients with certain forms of adult-onset blindness, the outermost photoreceptor layer is nearly or completely degenerated, whereas some fraction of neurons in the inner retina, particularly the ganglion cells whose axons make up the optic nerve, are spared to some extent. Such patients are potential candidates for a retinal prosthesis. However, before a prosthetic device can be optimally used, it is vital to understand how visual information is encoded in the pattern of electrical impulses traveling down the optic nerve.

The output of the retina, and indeed of most neurons in the central nervous system, cannot be classified in conventional electrical engineering terms as either analog or digital; rather, neuronal output consists of a temporal sequence of impulses, or spikes (see Figure 2). It is therefore vital to understand how visual information is encoded in spike trains traveling down the optic nerve. In the absence of stimulation, most ganglion cells fire spikes randomly at a background rate much lower than their maximum firing frequency. The conceptual diagram in Figure 2 depicts the spike trains from two clusters of neighboring ganglion cells. Clusters are indicated rather than single ganglion cells both to justify the high signal-to-noise ratio illustrated in the figure and to take into account the spatial convergence of optic-nerve fibers onto target neurons deep within the brain. When a cluster is stimulated by a small spot roughly equal in size to the excitatory portion of the receptive field (the local region of the visual

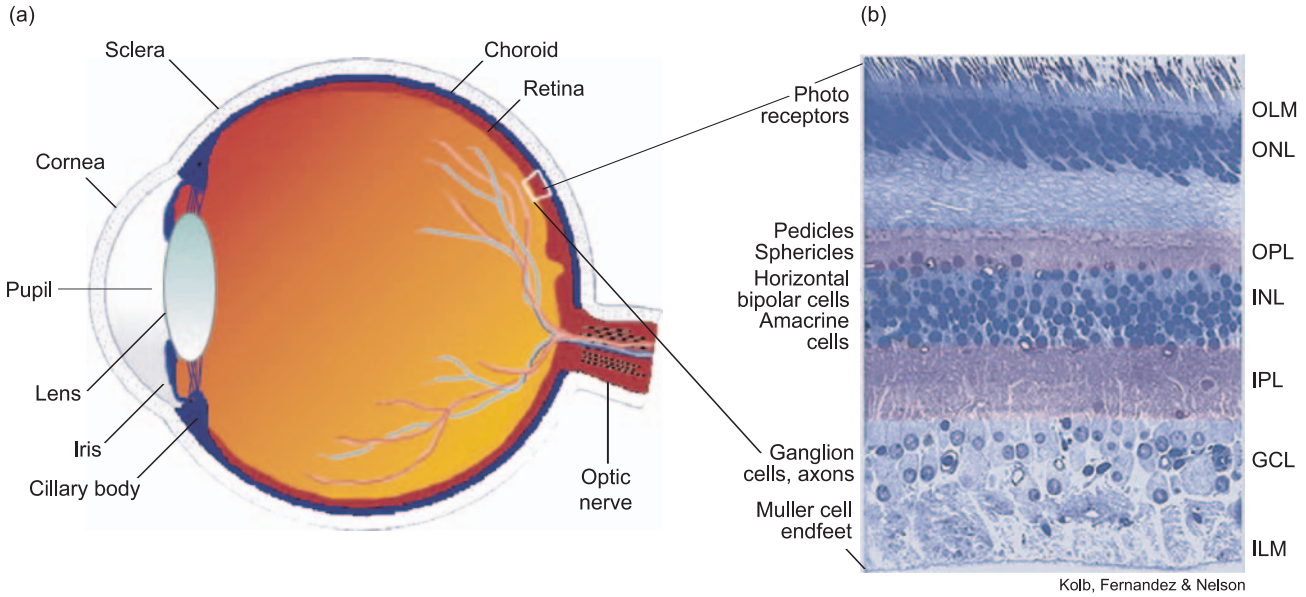


Figure 1. The Retina

(a) Located at the back of the eyeball, the retina consists of many types of neurons arranged in a layered structure. The light-sensitive cells (photoreceptors—rods and cones) are in the outermost layer, farthest from the incoming light. In front of the photoreceptors are neurons that perform specialized processing. At the innermost layer are ganglion cells whose axons make up the optic nerve. (b) A photomicrograph of a cross section of the retina reveals those distinct processing layers.

(Courtesy of Webvision http://webvision.med.utah.edu/ at the Moran Eye Center.)

space to which the indicated group of cells best respond), the firing rate, that is, the average number of spikes per time interval, increases markedly in proportion to the contrast between the spot's intensity and the light intensity immediately surrounding the spot. Regardless of the stimulus intensity, however, the timing of the individual spikes in response to a small spot remains more-or-less randomly distributed such that the spikes elicited by two small spots will typically be entirely uncorrelated. The observation that the mean firing rate is proportional to the stimulus intensity while the spikes themselves are distributed randomly in time is the basis for the rate-code hypothesis, which posits that information is transmitted only by the mean number of spikes per time interval irrespective of their precise timing.

There is evidence, however, that the rate-code hypothesis is incomplete. As the size of the two spots increases, the total number of spikes per time interval is reduced somewhat by lateral inhibition, but the most striking effect is the appearance of an oscillation in the firing rate, causing spikes to occur in relatively narrow clusters, or bursts. The phase of the underlying oscillation drifts randomly over time, so that the bursts evoked by separate spots will rapidly become uncorrelated even if both sets of neurons are modulated at a similar frequency. Remarkably, when the two groups of neurons are stimulated by a single large spot, the groups' underlying oscillations become strongly phase-locked, suggesting that the relative timing of spikes in the optic nerve can convey information about the spatial connections of features in the visual field. Such information is not conveyed by the local firing rate. To better illustrate the above encoding principles, it is useful to examine real physiological data. Figure 3 shows an intracellular recording from a retinal ganglion cell in response to a sinusoidally varying light intensity. In Figure 3(a), the intensity of the light is shown as a function of time. Conceptually, the recorded trace in Figure 3(b) can be divided into two parts: a subthreshold membrane potential, exhibiting an approximately sinusoidal modulation, and action potentials, or spikes, which are the large impulses riding on top of the subthreshold membrane potential. This potential is not available to the brain because it represents the analog component of the response that is most directly proportional to the incident light intensity. Only the spikes riding on top of this potential are transmitted through the optic nerve to relay nuclei within the brain. Because each spike is, to a first approximation, identical to every other spike, information can be conveyed only by the temporal sequence of the spikes. To reveal the information embedded in experimentally recorded spike trains, neuroscientists typically average over many stimulus trials. The response histogram, obtained by combining spike trains from many stimulus trials, shows that the average firing rate of the recorded ganglion cell is roughly

proportional to the applied light intensity, except for an approximately 90° phase advance, which reflects the fact that the cell is sensitive to the rate of change of the light intensity, and a negative cutoff due to the fact that the firing rate cannot drop below zero—see Figure 3(c). According to the rate-code hypothesis, the multitrial rate histogram fully characterizes the information conveyed by neural spike trains. (Of course, the brain cannot perform a multitrial average, but it is assumed that the brain can extract similar information in real time by combining low-pass filtering and information from many cells.)

About 10 years ago, Wolf Singer's laboratory in Germany reported that retinal neurons also use the relative timing of spikes to encode information about visual stimuli that is not conveyed by their local firing rates (Neuenschwander and Singer 1996, Neuenschwander et al. 1999). Unlike the previous example, which involved an intracellular recording from a single cell, the data from Singer's laboratory consists of spike trains from two retinal neurons recorded simultaneously in response to a sustained light stimulus, either two separate short bars or a single long bar (Figure 4). In both cases, the edges of the bar stimuli were positioned over the receptive centers of the recorded cells so that, locally, the stimulation was approximately the same regardless of whether the stimulus consisted of one or two bars. Another difference from the previous example is that the spike trains in the Singer experiment were recorded with extracellular electrodes, which do not permit examination of the corresponding membrane potentials. However, spike trains recorded immediately outside the cell can be analyzed for temporal structure both individually and by pairs.

Correlation functions constructed from the individual spike trains revealed that the outputs of both retinal neurons were modulated by periodic oscillations—see the small boxes in Figures 4(a) and 4(b). The oscillations elicited by separate light bars were not phase-locked, as indicated by the relative absence of correlations between the two spike trains when stimulated by separate bars—

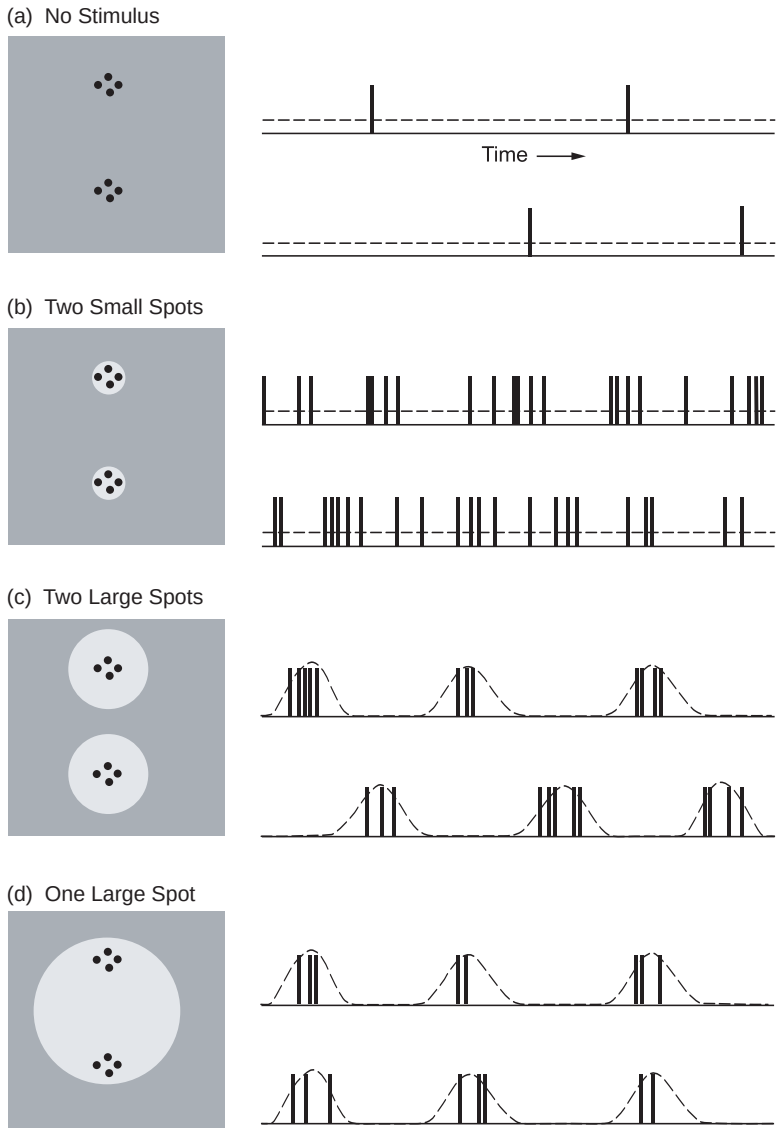


Figure 2. Retinal Responses to Light Stimuli

The gray boxes show two groups of retinal photoreceptors (black dots) exposed to spots of light during four experiments. At right are the outputs of the two groups of ganglion cells activated by the two groups of photoreceptors. (a) With no light stimulus, the outputs are random spikes. (b) When each group of photoreceptors is exposed to a small spot of light, the average firing rate of the overlying ganglion cells increases in proportion to the ratio of the intensity of the spot to the intensity of the region immediately surrounding the spot, but the spikes still occur more or less randomly. (c) As the spot size increases, the firing rate decreases somewhat, and the spikes bunch up. However, the bunches are not synchronized. (d) When the spot size is large enough to cover both groups of photoreceptors, the bunches become synchronized.

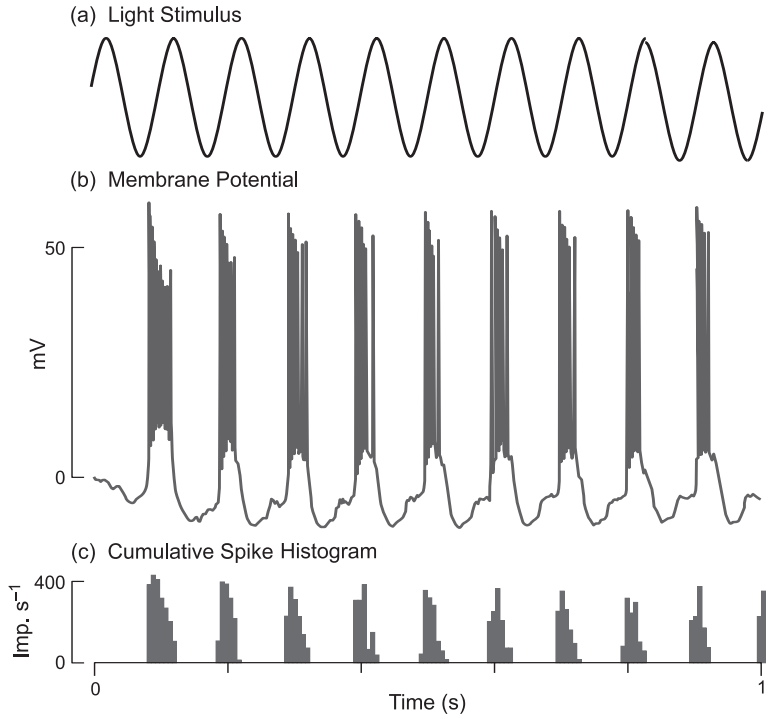


Figure 3. Responses of a Single Retinal Ganglion Cell

A motion-sensitive ganglion cell from a mammalian retina was stimulated by a spot of light whose intensity varied sinusoidally in time. (a) The intensity of the light spot is shown as a function of time. (b) The subthreshold membrane potential recorded at the soma (cell body) consists of action potentials, or spikes, riding on top of an approximately sinusoidal modulation. Only the spikes are transmitted to relay nuclei in the brain. (c) A response histogram is constructed from spikes accumulated over many identical stimulus trials, showing that the average firing rate is approximately proportional to the applied stimulus intensity except for a phase shift and a lower cutoff at 0 Hz. (Dacey and Lee 1994. Reprinted with permission from *Nature*.)

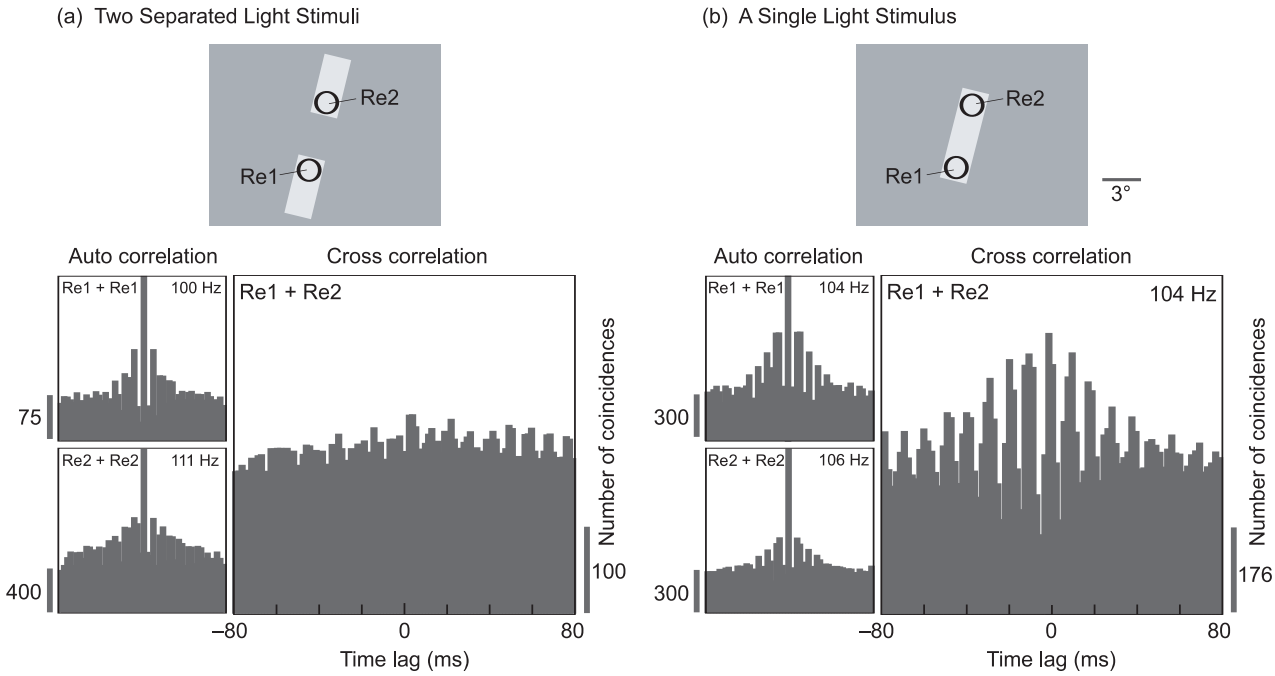


Figure 4. Responses of Two Retinal Ganglion Cells

Two cat-retina ganglion cells separated by 6 degrees are monitored by electrodes Re1 and Re2, respectively. Spike trains generated in response to rectangular light stimuli were recorded simultaneously from each electrode, and autocorrelation and cross-correlation histograms were computed. (a) When two distinct rectangular light stimuli are used, strong oscillations are present in the autocorrelation histogram of each ganglion cell, but the cross-correlation histogram between the two ganglion cells is essentially flat. The sharp peaks in the autocorrelation histograms correspond to the spike bunching in Figures 2(c) and 2(d). (b) When a single rectangular stimulus that connects the region between the cells is used, the oscillations in firing rate of the two ganglion cells become strongly phase locked, generating a strong oscillation in the cross-correlation histogram. The cross-correlation oscillation corresponds to the synchronization of the spike bunches in Figure 2(d). (Neuenschwander 1996. This figure was redrawn courtesy of *Nature*.)

see the large box in Figure 4(a). On the other hand, the evoked oscillations were tightly phase-locked when activated by a single bar, as indicated by the strong periodic modulations in the corresponding correlation function illustrated in the large box in Figure 4(b). Thus, the timing of retinal ganglion-cell spikes, especially with respect to the phase differences between the oscillatory responses of separate groups of ganglion cells, can convey information relevant to the spatial separation or spatial binding of visual features.

A Computer Model

We constructed a computer model of the retina (Figure 5) to study how temporal codes in the retina might be generated and what types of visual information such codes might convey. Very few circuits in the central nervous system have been completely characterized, and the circuits of the retina are no exception. However, enough is known about retinal anatomy and physiology to allow the construction of a model that accounts for many aspects of experimentally recorded light responses in a manner consistent with general patterns of neuronal connectivity found in most vertebrate species. Moreover, by requiring the model to account for a wide range of experimental data, we were able to infer some aspects of the unknown anatomy and physiology. In particular, by increasing the range of observed phenomena explained by the model, we were able to eliminate alternative implementations that were inconsistent with published recordings.

The model consisted of five distinct cell types illustrated in Figure 5(a): bipolar cells, three classes of amacrine cells, corresponding to the small, large, and polyaxonal subtypes, and ganglion cells. Bipolar cells are

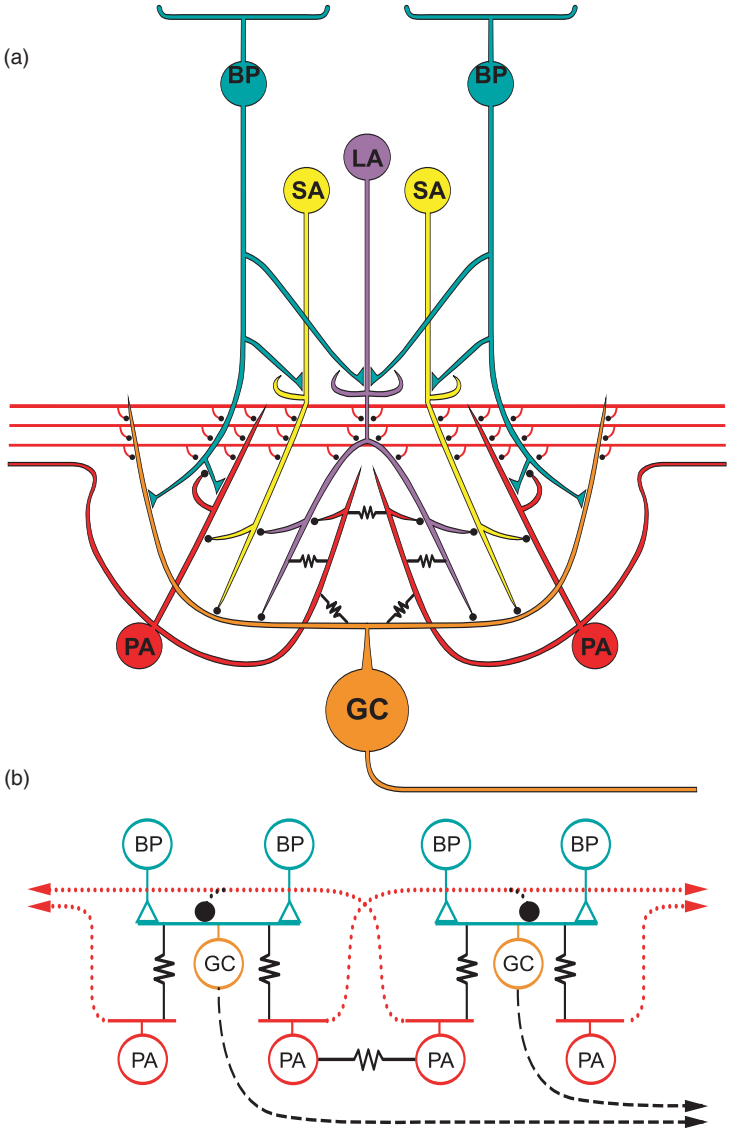


Figure 5. Computer Model of the Retina

(a) Our computer model consisted of five cell types: bipolar (BP) cells, small (SA), large (LA), and polyaxonal (PA) amacrine cells, and alpha ganglion (GC) cells, arranged in a 32 × 32 square mosaic with wrap-around boundary conditions. Although this side view of one of the mosaic's units shows only two BPs, there were actually four BPs. Light stimuli were simulated by injecting currents directly into the BPs. (Photoreceptors were not included in the model.) The inhibitory connections can be organized into three categories: *Feedforward and feedback inhibition.* Excitatory synapses from BPs were balanced by a combination of reciprocal synapses and direct inhibition of the GCs, mediated by the nonspiking amacrine-cell types. *Serial inhibition.* The three amacrine-cell types regulated each other through negative feedback loops. *Resonance circuit.* The PAs were excited locally through electrical synapses with GCs, and their axons gave rise to widely distributed inhibition that contacted all cell types, but most strongly the GCs and other PAs. Not all connections present in the model are shown. (b) A simplified schematic diagram of the computer model shows how a combination of local excitation (triangles) carried by gap junctions (resistors) and long-range inhibition (empty circles) carried through axon-bearing amacrine cells (orange dotted lines and filled black circles) produced physiologically realistic oscillations dependent on stimulus size.

(Reprinted with the permission of Cambridge University Press.)

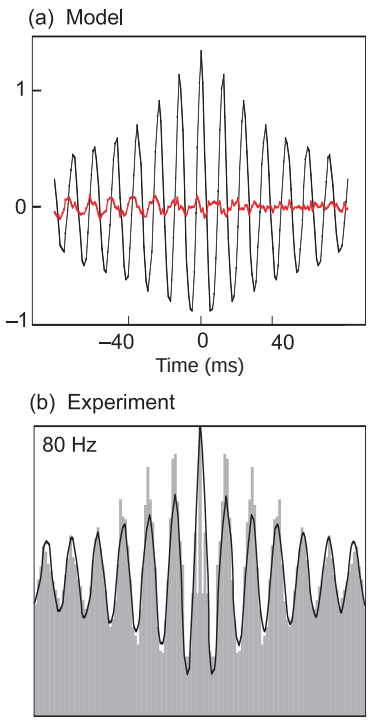


Figure 6. Light-Evoked Oscillations in the Retinal Model
(a) The average correlation function of the retinal model's output for pairs of ganglion cells exhibits an oscillation whose amplitude, frequency, and duration are similar to those of (b), the correlation function for experimentally recorded spike trains from cat retina cells in response to an analogous stimulus.

relay neurons that receive synaptic input from photoreceptors (not shown) and provide the principal excitatory drive to the other neuron types. The bipolar cells are therefore critical elements in the "vertical" pathway representing the main direction of information flow from the photoreceptors to the optic nerve. On the other hand, amacrine cells mediate lateral interactions that are essential for processing and encoding visual signals. Indeed, without amacrine-cell interactions, corresponding to the "horizontal" pathway, the output of the retina would simply replicate the activity across the photoreceptor array. Nearly 30 different kinds of amacrine cells have been described, and it would not have been possible to include them all in the model. Instead, we included only the minimum number of amacrine-cell types necessary to account for the basic features of retinal light responses, as well as for the synchronous oscillations evoked by large stimuli.

While the circuitry incorporated into the model is somewhat complicated, the connections can be grouped into three major categories:

Excitation. The bipolar cells, which relayed visual signals from the photoreceptors to all the other cell types, were the only source of excitation in the model.

Local inhibition. The model amacrine cells made feedforward inhibitory synapses onto the ganglion cells and feedback inhibitory synapses onto the bipolar cells and serially inhibited each other. These local inhibitory synapses acted to increase the dynamical range of the model retina, by negative feedback, and further contributed to shaping light-evoked activity so as to amplify the responses to both spatial and temporal contrast.

Long-range axon-mediated feedback. The polyaxonal amacrine cells received local excitation from ganglion cells by electrical synapses or gap junctions and, in turn, made long-range inhibitory connections onto all cell types. This delayed negative feedback circuit accounted for the generation of oscillatory responses in the model retina and has been redrawn in Figure 5(b) to emphasize the circuit's essential components and their interconnections. Further details of the model, particularly its ability to account for experimental data as well as its numerical stability and robustness to parameter variation, are available elsewhere (Kenyon et al. 2004a, Kenyon et al. 2004b, Kenyon et al. 2003a, Kenyon et al. 2003b).

Oscillations

To assess the stimulus-evoked oscillations in the retinal model, correlations were computed between the spike trains arising from all pairs of ganglion cells activated by a large spot, and the results were combined into an averaged correlation measure—refer to Figure 6(a). The amplitude, frequency, and persistence of the periodic modulations in the averaged correlation function obtained from the retinal model were qualitatively similar to those observed experimentally in recordings of cat ganglion cells responding to large, high-contrast spots, as illustrated in Figure 6(b). For both the cat retina and the retinal model, the correlation amplitude falls off with increasing delay, eventually returning to approximately baseline levels after several cycles of the underlying oscillation. In both sets of data, the phases of the underlying oscillations drift randomly over time, so that firing activity becomes uncorrelated over sufficiently long delays. This time drift is a fundamentally nonlinear phenomenon arising from the threshold nature of spike generation. In contrast, the phase of a linear harmonic oscillator is always fixed relative to the stimulus onset. The retinal model thus captures an essential nonlinear property of the biological circuitry. Moreover, the good qualitative agreement between theory and experiment implies that the free parameters in the model, particularly those involving the axon-mediated feedback circuit, were likely to be reasonably close to their true physiological values. Other comparisons with physiological data were used to verify additional aspects of the model.

The retinal model was also able to account for the experimentally observed size dependence of retinal oscillations (Figure 7). Specifically, the oscillations evoked by stimuli of various sizes in our retinal model

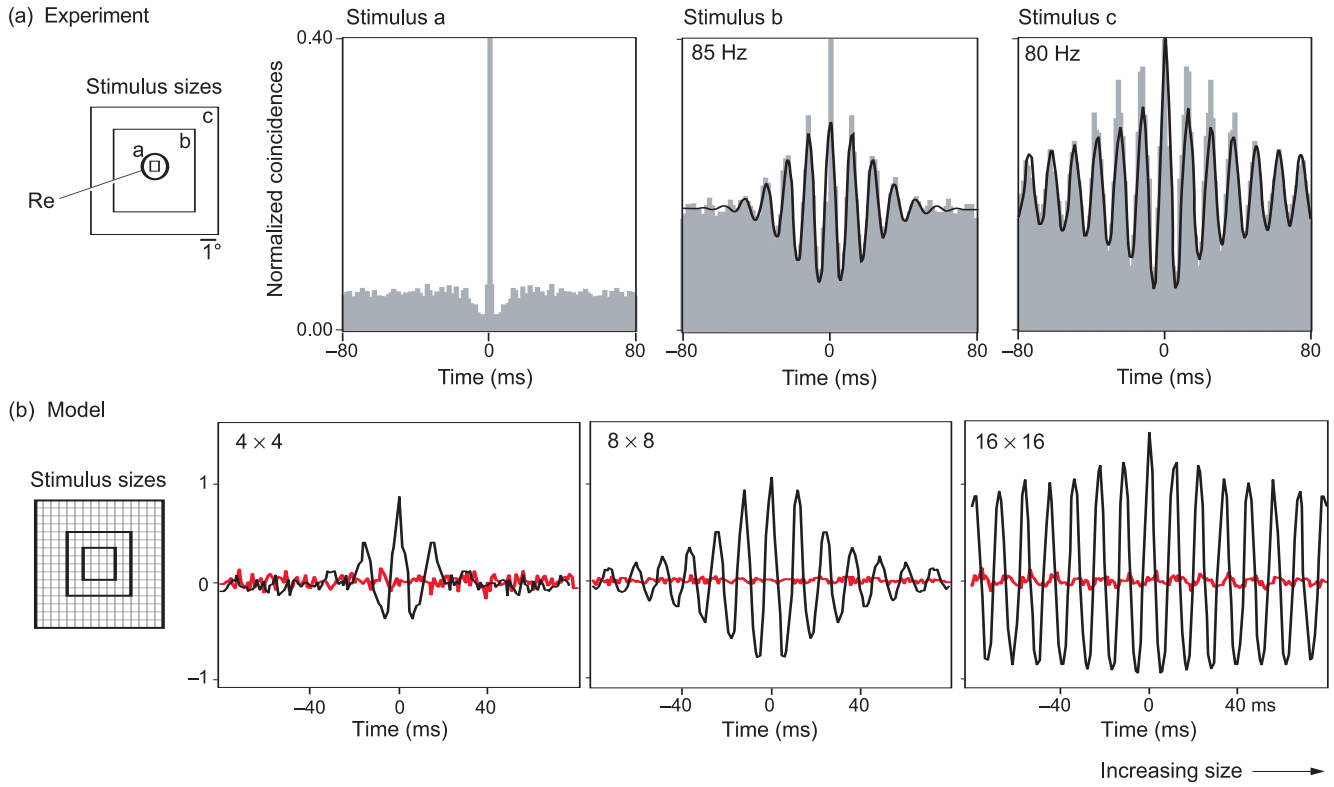


Figure 7. Stimulus-Size Dependence of Retinal Oscillations
(a) The correlation function computed between experimentally recorded spike trains from cat retina cells exhibits a strong increase in oscillatory activity with increasing stimulus size. (b) The correlation function computed for the oscillations produced by our retinal model exhibits a similar size dependence. (Experimental data from Neuenschwander (1996). Redrawn courtesy of *Nature*.)

were similar to those measured from the cat retina. In both sets of data, small stimuli evoked little or no oscillatory response, whereas large stimuli evoked oscillations with very large amplitudes. Because the axon-mediated feedback was spread out over a wide retinal area, only large stimuli could evoke strong oscillatory responses. The notion that oscillations are associated with large stimuli led us to put forward a novel hypothesis about the types of visual signals encoded in the periodic temporal structure of retinal spike trains. We discuss this hypothesis in more detail below.

The model also accounted for a high-frequency resonance observed in the responses of certain retinal neurons to temporally modulated stimuli (Figure 8). The temporal modulation transfer function (tMTF) measures how strongly the output of a system is modulated as a function of the frequency of a sinusoidal input. Harmonic or oscillatory systems typically exhibit a resonance frequency, at which the output of the system can be driven to relatively large amplitudes. The presence of high-frequency oscillations in retinal light responses suggests that there will be a corresponding resonance in the tMTF, given by the amplitude of the sinusoidal modulation in the ganglion-cell firing rate when plotted as a function of the frequency of the applied stimulus. As expected, both the cat and model retinas exhibit sharp resonance peaks in their tMTFs at frequencies above 60 hertz, at which frequency value oscillatory responses are also observed. Moreover, the model accounted not only for the resonance itself but also for the associated kink in the phase-response curve, which plots how much the phase of the output modulation is retarded or advanced relative to the sinusoidal input. Such kinks are not present in the phase-response curves of simple harmonic oscillators. Using our computer model, we were able to show that the kink in the phase-response curve obtained from retinal ganglion cells was due to entrainment. When driven at relatively low modulation frequencies, the oscillations produced by retinal circuitry, whose phase drifts randomly over time, quickly become independent of the phase of the driving stimulus. As the frequency of the driving stimulus approaches the resonance frequency, however, the two oscillations become entrained, causing an abrupt advance in the phase-response curve. Such resonances may

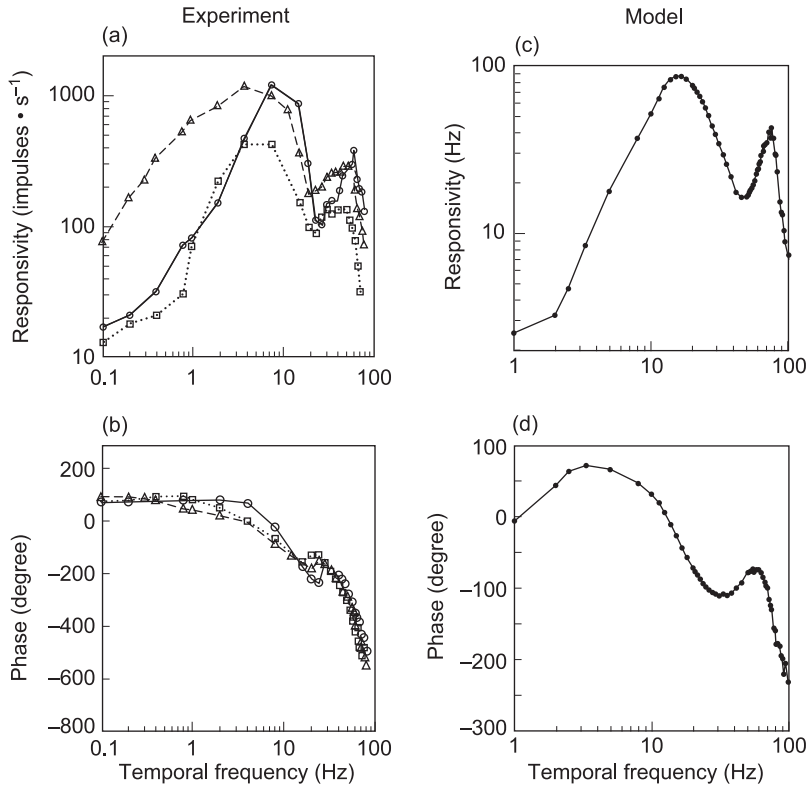


Figure 8. Temporal Modulation Transfer Functions (tMTFs)
(a) The tMTF recorded from cat-retina ganglion cells is obtained by plotting the magnitude of the fundamental Fourier component in the response histogram as a function of the temporal modulation frequency of the applied stimulus. The maximum response occurs at a broad low-frequency resonance, between 10 Hz and 20 Hz, but there is also a prominent high-frequency resonance at around 70 Hz. (b) The phase-response curve corresponding to the data shown in (a) is plotted as a function of temporal modulation frequency. The phase-response curve exhibits a prominent kink at frequencies near the rising phase of the resonance peak. (c) A similar resonance is present in the tMTF recorded from ganglion cells in the retinal model in response to a temporally modulated spot. (d) Likewise, the phase-response curve of the model ganglion cells also exhibits a kink near the onset of the resonance peak. The kink is caused by entrainment of the retinal oscillations by the applied stimulus.
(Reproduced from the *Journal of General Physiology*, 1987, Vol. 89, pp. 599–628, by copyright permission of The Rockefeller University Press.)

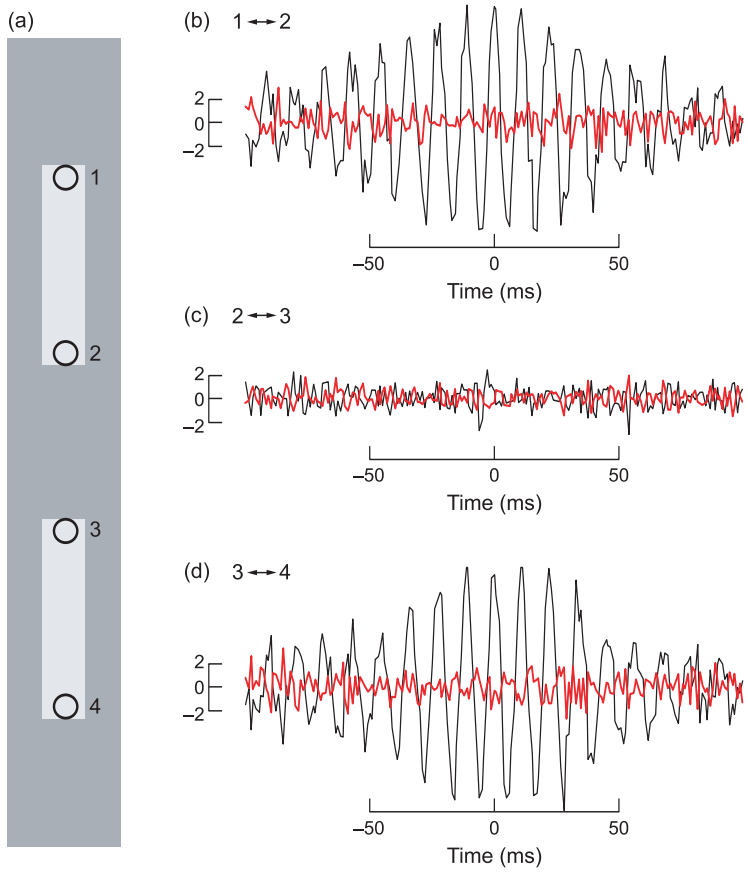


Figure 9. Stimulus-Selective Oscillations
(a) Two bar-shaped light stimuli are shown in relation to the receptive field centers of four simultaneously recorded ganglion cells. Cross-correlation histograms were computed during the plateau portion of the responses for pairs of ganglion cells at opposite ends of the same bar or at opposing tips of separate bars. All ganglion cell pairs were separated by 7 diameters. The cross-correlation histograms were computed for pair 1,2 from the upper bar (b), pair 2,3 from the two separate bars (c), and pair 3,4 from the lower bar (d). The histograms exhibit significant oscillations only for pairs stimulated by the same bar.
(Courtesy of Wolf Singer and colleagues.)

also be relevant to the effective operation of a retinal prosthesis by enabling strategies for selectively activating certain types of retinal neurons at the characteristic frequencies of stimulation.

Finally, the retinal model was able to reproduce the stimulus selectivity of retinal oscillations first reported by Wolf Singer's laboratory, as outlined above. By examining the relative timing of spikes produced by retinal ganglion cells responding to either the same or to different objects, we were able to show that model elements activated by the same large object were strongly correlated, or phase locked, by a common underlying oscillation at a frequency of approximately 100 hertz (see Figure 9). Pairs of model retinal neurons activated by different objects, however, were not correlated; that is, the phases of their underlying oscillations varied randomly with respect to each other. Thus, our retinal model captures the interesting property of biological neurons that their evoked oscillations in responses to large visual features are stimulus specific and are only phase locked for cells responding to the same contiguous object. The above results illustrate how the retinal model was able to account for many of the main experimentally observed aspects of oscillatory phenomena.

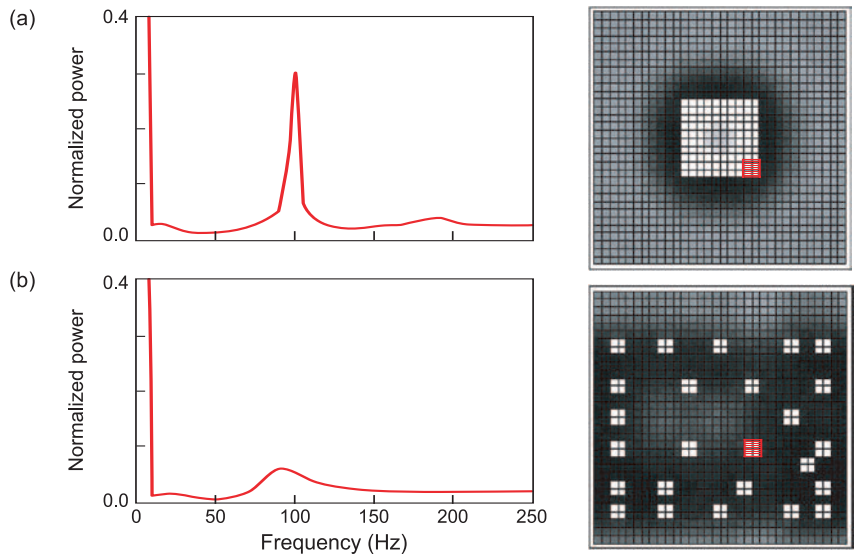


Figure 10. See Globally, Spike Locally
We exposed 25 clusters of ganglion cells in our computer-model array of 32 × 32 cells to two different high-contrast light stimuli and computed the power spectrum from the output of a single cluster (shown in red) for each exposure condition. Each cluster consisted of 2 × 2 neighboring ganglion cells. The power spectra (left) were normalized by the total average firing rate for each exposure condition (right). (a) For a cluster that was part of a large illuminated area, the power spectrum peaked sharply between 60 Hz and 120 Hz. (b) For a cluster illuminated in isolation, the power spectrum exhibited only a small hump.

What Do Oscillations Encode?

Having established the biological plausibility of the retinal model, we then used computer simulations to explore what information stimulus-specific oscillations might convey to the brain. Because our model is consistent with the known anatomy and physiology of the cat retina, the model can provide a useful tool for investigating how information can be encoded in the temporal structure of spike trains propagating down the optic nerve. Based on the stimulus selectivity of retinal oscillations, as well as their observed size dependence, a collection of disconnected spots will elicit only weak periodic modulations in optic-nerve fibers, whereas a single large stimulus will elicit strong period modulations. To test this idea, we exposed our computer-model array of 32 × 32 ganglion cells to two different light stimuli. The first stimulus was a large, square spot covering 25 contiguous clusters of cells—refer to Figure 10(a). Each cluster consisted of 2 × 2 cells. For this stimulus, the power spectrum of the spike trains from a single cluster exhibited a large, sharp peak at around 100 hertz. However, when the array was exposed to 25 small, isolated spots, each of which covered exactly one cluster but otherwise elicited approximately the same total number of spikes per time interval, there was only a small hump in the power spectrum, as shown in Figure 10(b). These results suggest that the periodic temporal structure in the spike trains obtained from small clusters of neighboring neurons encodes the overall size of the object to which the clusters respond.

To investigate the above hypothesis, we were guided by two principles: (1) Because it takes us only a fraction of a second to form a visual impression, the information conveyed by stimulus-specific oscillations must be available on short, physiologically meaningful time scales—roughly a few hundred milliseconds. (2) Because the spatial convergence of retinal neurons onto target cells in the brain is rather low, with each target cell receiving input from only a few retinal neurons, the

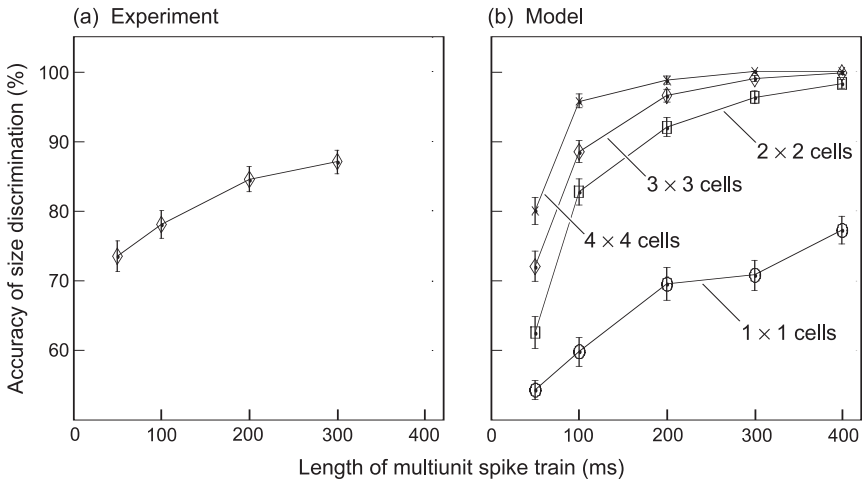


Figure 11. Size Discrimination

A Bayes discriminator was used to classify light spots as either "smaller" or "larger" from the single-trial oscillatory activity of (a) cat retina ganglia or (b) the ganglia in our retinal computer model. For the multiunit spike trains recorded from cat retina ganglia, the percentage of correctly classified trials ranged from ~73% to ~87% as the length of the multiunit spike train segment increased from 50 to 300 ms. The percentage of correctly classified trials using multiunit spike trains from the retinal model improved with longer analysis windows and as more ganglion cells were included in the spike record.

information conveyed by stimulus-specific oscillations must be available locally in the firing activity of a similar number of neighboring cells. We therefore used the retinal model to quantify the information conveyed about the global properties of a stimulus, in this case the total size of the object, by a 2 × 2 neighborhood of retinal output neurons in a few hundred milliseconds. At the same time, having received data from Wolf Singer's laboratory recorded from output neurons in the cat retina under analogous experimental conditions, we were able to test directly the predictions of our retinal model.

One of our studies tested our ability to determine if a group of neighboring cells was responding to a small or a large object from the group's local firing activity alone. In Figure 11, we plot the results of this study in terms of the "accuracy of size discrimination," which was equal to the fraction of trials in which the total size of the stimulus could be correctly inferred from the local firing activity. Random events were added to the model spike trains to ensure that the firing rate did not change as a function of stimulus size. The only cue available from the local firing activity regarding the total size of the stimulus was therefore the amplitude of the synchronous oscillations. Our results showed that, in model spike trains 300 milliseconds long, using as few as four spike trains from a small neighborhood (2 × 2 cells), nearly perfect accuracy can be achieved. The accuracy for the experimentally recorded spike trains was slightly lower, possibly reflecting to some extent the suboptimal recording conditions in which several different ganglion cell types contributed to the multiunit response. Overall, our modeling results imply that there is a tradeoff between the number of cells included in the analysis and the total time allowed for accomplishing the size-discrimination task. Specifically, as more cells were included in the analysis, shorter periods were required to achieve the same accuracy.

Why would it be important for retinal neurons to convey information about stimulus size in their local firing activity? Studies of a frog's retina may provide the answer. Tachibanna's laboratory in Japan has shown that the frog retina has specialized neurons, called dimming detectors, that exhibit strong synchronous oscillations when activated by a large dimming object but do not exhibit such oscillations when activated by a small dimming object (Ishikane et al. 1999). To a frog, a small dimming spot could be a fly or other food source, but a large dimming spot is more likely to be a bird or other dangerous predator. In this situation, one can easily appreciate why size matters. ■

Acknowledgments

We gratefully acknowledge the following collaborators in undertaking the described studies: Greg Stephens, Mark Flynn, and Benjamin Barrowe.

Further Reading

Frishman, L. J., A. W. Freeman, J. B. Troy, D. E. Schweitzer-Tong, and C. Enroth-Cugell. 1987. Spatiotemporal Frequency Responses of Cat Retinal Ganglion Cells. *J. Gen. Physiol.* **89**: 599.

Geddes, L. A., and L. E. Baker. 1967. The Specific Resistance of Biological Material—A Compendium of Data for the Biomedical Engineer and Physiologist. *Med. Biol. Eng.* **5** (3): 271.

Humayun, M. S., J. D. Weiland, G. Y. Fujii, R. Greenberg, R. Williamson, J. Little, et al. 2003. Visual Perception in a Blind Subject with a Chronic Microelectronic Retinal Prosthesis. *Vision Res.* **43**: 2573.

Ishikane, H., A. Kawana, and M. Tachibana. 1999. Short- and Long-Range Synchronous Activities in Dimming Detectors of the Frog Retina. *Vis. Neurosci.* **16** (6): 1001.

Jensen, R. J., J. F. Rizzo III, O. R. Ziv, A. Grumet, and J. Wyatt. 2003. Thresholds for Activation of Rabbit Retinal Ganglion Cells with an Ultrafine, Extracellular Microelectrode. *Invest. Ophthalmol. Vis. Sci.* **44** (8): 3533.

Kenyon, G. T., J. Theiler, J. S. George, B. J. Travis, and D. W. Marshak. 2004. Correlated Firing Improves Stimulus Discrimination in a Retina Model. *Neural Comput.* **16** (11): 2261.

Kenyon, G. T., J. Theiler, D. W. Marshak, B. Moore, J. Jeffs, and B. J. Travis. 2003. Firing Correlations Improve Detection of Moving Bars. In *Proc. Int. Joint Conf. Neural Network.* **1–4**: 1274.

Kenyon, G. T., B. J. Travis, J. Theiler, J. S. George, G. J. Stephens, and D. W. Marshak. 2004. Stimulus-Specific Oscillations in a Retinal Model. *IEEE Trans. Neural Network.* **15** (5): 1083.

Kenyon, G. T., B. Moore, J. Jeffs, K. S. Denning, G. J. Stephens, B. J. Travis et al. 2003. A Model of High-Frequency Oscillatory Potentials in Retinal Ganglion Cells. *Vis. Neurosci.* **20** (5): 465.

Margalit, E., M. Maia, J. D. Weiland, R. J. Greenberg, G. Y. Fujii, G. Torres et al. 2002. Retinal Prosthesis for the Blind. *Surv. Ophthalmol.* **47**(4): 335.

Neuenschwander, S., M. Castelo-Branco, and W. Singer. 1999. Synchronous Oscillations in the Cat Retina. *Vision Res.* **39**: 2485.

Neuenschwander, S., and W. Singer. 1996. Long-Range Synchronization of Oscillatory Light Responses in the Cat Retina and Lateral Geniculate Nucleus. *Nature* **379**: 728.

Travis, B. J., and A. D. Chave. 1989. A Moving Finite Element Method for Magnetotelluric Modeling. *Phys. Earth Planet. Inter.* **53**: 432.

Unsworth, M. J., B. J. Travis, and A. D. Chave. 1993. Electromagnetic Induction by a Finite Electric Dipole Source over a 2-D Earth. *Geophys.* **58** (2): 198.

For further information, contact Garrett Kenyon (505) 667-1900 (gkenyon@lanl.gov).

Modeling Stimulation by a Retinal Prosthesis

We have begun developing a three-dimensional (3-D) model of the retinal extracellular space. Existing software, originally developed for modeling the flow of ground water through porous material [(Unsworth et al. 1989), (Travis and Chave 1989)], has been adapted to calculate the potential gradients produced by an arbitrary distribution of stimulating electrodes. The flow of water in porous media (that is, sedimentary rock) and the flow of current through the extracellular space are mathematically identical problems, allowing powerful software tools developed in one context to be applied to the other. Our basic strategy is to avoid the fully interacting problem that requires solving for the intercellular and extracellular potentials simultaneously. Instead, we take advantage of the fact that the ephaptic, or incidental coupling between retinal neurons via the extracellular space is small and that any significant extracellular potential gradients will be due almost entirely to external stimulation. Prosthetic stimulation can therefore be modeled in two distinct steps: (1) calculate the extracellular potentials produced by the applied currents and (2) compute how the resulting gradients act upon dendritic and axonal processes within the retina. In principle, the same technology could be used in reverse; the normal light-evoked responses of retinal neurons could be calculated before hand and the resulting membrane currents could be used as sources to estimate local field potentials. Such technology could eventually be useful for connecting realistic simulations of retinal circuits to clinical measures such as the electroretinogram.

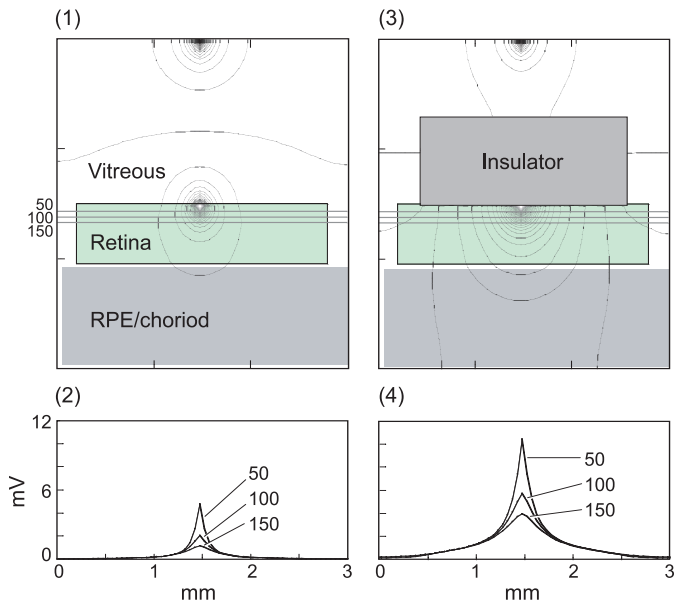


Figure A. The Influence of Both Anatomical and Non-Anatomical Structures of the Distribution of Extracellular Currents
(1) This contour plot is of extracellular potentials due to dipole stimulation of the retina and associated structures (see labels). (2) Profiles of extracellular potentials are shown at three different depths, 50, 100, and 150 μ. Panels (3) and (4) show the same organization as (1) and (2) with the addition of an insulating block representing the prosthetic multielectrode array itself.

As preliminary data, we have developed a simplified model of the retina and associated structures in which the various anatomical elements, consisting of the vitreous, retina and retina pigment epithelium (RPE)/choroid, as well as the multielectrode array itself, were represented as rectangular blocks (Figure A). The bulk conductivity of each element was based on published values (Geddes and Baker 1967), with the bulk conductivity of the multielectrode array set to zero. The computer model was used to calculate the extracellular potential gradients produced by a 1 microampere anodic current pulse passed through a single stimulating electrode, 10 microns in diameter, located on the vitreous surface. The return electrode was placed in the vitreous cavity 400 microns away along a line perpendicular to the retinal surface. It is well established that cathodic current pulses are much more effective for stimulating retinal neurons, but for modeling the spatial distribution of extracellular currents, the overall sign is irrelevant.

In the absence of an insulating barrier above the retina, transverse bipolar stimulation produced a dipole field that was nearly mirror symmetric, with the slight deviations arising from the conductivity differences of the various tissue components—Figure A(1). The spatial profiles of the extracellular potentials parallel to the retinal surface were examined as a function of depth from the stimulating electrode—Figure A(2). At a depth of 50 microns, the maximum value of the extracellular potential directly underneath the electrode was just under 6 millivolts,

and fell off laterally with a length constant on the order of 100 microns. Deeper in the retina, the extracellular potential fell off more gradually as the radial component away from the electrode became smaller in the lateral direction.

A prosthetic device would not consist of a single pair of electrodes, however, but of a multielectrode array contained in an insulating package. We therefore used the computer model to examine how a large insulator affixed to the vitreous surface would affect current flow within the retina—Figures A(3) and A(4). Our results show that by forcing more of the current into the retina, a large non-conducting barrier can substantially enhance the effects of prosthetic stimulation. Inserting a representation of the prosthetic device into the 3-D model approximately doubled the extracellular potential gradients produced by the same 1microampere current pulse applied previously. These results illustrate the general principle of how the 3-D geometry of the retina and associated structures, as well design of the prosthetic device itself, can have a large impact on the spatial distribution of externally applied currents.

Computer models can also be used to investigate how cellular properties, such as dendritic morphology and orientation, influence responses to prosthetic stimulation. As a preliminary step, we examined the changes in membrane potential produced by a 1-microampere cathodic current as a function of orientation, either vertical or horizontal—Figure B. A 50-micron passive segment, representing a bipolar cell axon or ganglion cell dendrite, was centered 75 microns from the vitreous surface directly underneath the stimulating electrode. When the passive cable segment was oriented vertically, as would likely be the case for bipolar cell axons, there was approximately a 1 millivolt depolarization at the proximal tip closest to the electrode. On the other hand, when the same passive segment was oriented horizontally, as would be predominantly the case for ON ganglion cell dendrites, there was virtually no change in the membrane potential.

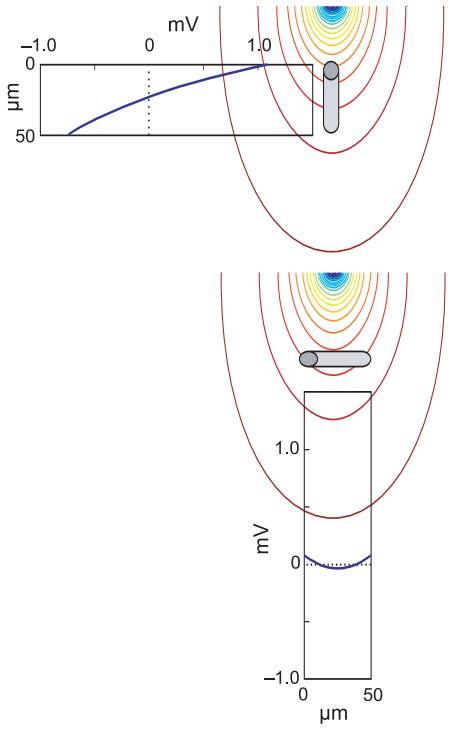


Figure B. Effects of Orientation
Shown here is the change in membrane potential along a 50-μ passive cable in response to a 1-μA current. A vertically oriented segment experiences a maximum depolarization of nearly 1 mV at its proximal tip, while a horizontally oriented segment is virtually unaffected by the stimulus. Cable centers were 75 μ from the electrode.

The strong influence of orientation is a direct consequence of the fact that neural processes are activated by gradients in the extracellular potential and are insensitive to the average magnitude, or constant offset. For transverse stimulation in which the dipole axis is perpendicular to the retinal surface, vertically oriented processes lay across equipotential contour lines and thus along the maximum gradient. Horizontally oriented processes, on the other hand, lie mostly parallel to equipotential contour lines and thus experience minimal gradients along their length. The dendrites of ON ganglion cells tend to be oriented laterally and are therefore likely to have higher activation thresholds than vertically oriented processes. Ganglion cell axons are also oriented in a predominantly lateral direction and thus are less strongly activated by transverse currents, but this effect is potentially countered by their closer proximity to the electrode array.

Finally, if an anodic current pulse had been applied instead, the proximal tip of the vertically oriented segment would have been hyperpolarized rather than depolarized. It has been reported that bipolar cell activation thresholds are lower for transverse cathodic currents than for equivalent anodic currents (Jensen et al. 2003). Our simulations provide insight into this phenomenon. The passive cable segments used in our preliminary study were electronically short and thus can be treated as approximately isopotential. The extracellular potential, on the other hand, grows more negative in the direction of the cathode, here assumed to be on the vitreous surface. At the proximal tip of a vertically oriented process, the extracellular potential will be most negative and thus closest to the intracellular potential (assumed to be uniform), causing the membrane potential at that point to be depolarized from the resting potential. At the distal tip, on the other hand, the difference between the intracellular and extracellular potential is maximal, and the corresponding membrane potential is hyperpolarized. This example shows how the stimulation protocol must be designed in tandem with information about cellular morphology.

Leonardo da Vinci's illustration of the swirling flow of turbulence.
(*The Royal Collection © 2004, Her Majesty Queen Elizabeth II*)

The Turbulence Problem

An Experimentalist's Perspective

Robert Ecke

Turbulent fluid flow is a complex, nonlinear multiscale phenomenon, which poses some of the most difficult and fundamental problems in classical physics. It is also of tremendous practical importance in making predictions—for example, about heat transfer in nuclear reactors, drag in oil pipelines, the weather, and the circulation of the atmosphere and the oceans. But what is turbulence? Why is it so difficult to understand, to model, or even to approximate with confidence? And what kinds of solutions can we expect to obtain? This brief survey starts with a short history and then introduces both the modern search for universal statistical properties and the new engineering models for computing turbulent flows. It highlights the application of modern supercomputers in simulating the multiscale velocity field of turbulence and the use of computerized data acquisition systems to follow the trajectories of individual fluid parcels in a turbulent flow. Finally, it suggests that these tools, combined with a resurgence in theoretical research, may lead to a "solution" of the turbulence problem.

Many generations of scientists have struggled valiantly to understand both the physical essence and the mathematical structure of turbulent fluid motion (McComb 1990, Frisch 1995, Lesieur 1997). Leonardo da Vinci (refer to Richter 1970), who in 1507 named the phenomenon he observed in swirling flow "la turbolenza" (see the drawing on the opening page), described the following picture: "Observe the motion of the surface of the water, which resembles that of hair, which has two motions, of which one is caused by the weight of the hair, the other by the direction of the curls; thus the water has eddying motions, one part of which is due to the principal current, the other to the random and reverse motion."

Two aspects of da Vinci's observations remain with us today. First, his separation of the flow into a mean and a fluctuating part anticipates by almost 400 years the approach taken by Osborne Reynolds (1894). The "Reynolds decomposition" of the fluid velocity into mean and fluctuating parts underpins many engineering models of turbulence in use today.[1] Second, da Vinci's identification of "eddies" as intrinsic elements in turbulent motion has a modern counterpart: Scientists today are actively investigating the role of such structures as the possible "sinews" of turbulent dynamics.

Long after da Vinci's insightful observations, a major step in the description of fluid flows was the development of the basic dynamical equation of fluid motion. The Euler equation of motion (written down in the 18th century) describes the conservation of momentum for a fluid without viscosity, whereas the Navier-Stokes equation (19th century) describes the rate of change of momentum at each point in a viscous fluid. The Navier-Stokes equation for a fluid with constant density ρ and constant kinematic viscosity ν is

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{\nabla P}{\rho} + \nu \nabla^2 \mathbf{u} \quad , \qquad (1)$$

with $\nabla \cdot \mathbf{u} = 0$, which is a statement of fluid incompressibility and with suitable conditions imposed at the boundaries of the flow. The variable $\mathbf{u}(\mathbf{x},t)$ is the (incompressible) fluid velocity field, and $P(\mathbf{x},t)$ is the pressure field determined by the preservation of incompressibility. This equation (when multiplied by ρ to get force per unit volume) is simply Newton's law for a fluid: Force equals mass times acceleration. The left side of Equation (1) is the acceleration of the fluid,[2] and the right side is the sum of the forces per unit mass on a unit volume of the fluid:[3] the pressure force and the viscous force arising from momentum diffusion through molecular collisions. Remarkably, a simple equation representing a simple physical concept describes enormously complex phenomena.

The Navier-Stokes equations are deterministic in the sense that, once the initial flow and the boundary conditions are specified, the evolution of the state is completely determined, at least in principle. The nonlinear term in Equation (1), $\mathbf{u} \cdot \nabla \mathbf{u}$, describes the advective transport of fluid momentum. Solutions of the nonlinear Navier-Stokes equations may depend sensitively on the initial conditions so that, after a short time, two realizations of the flow with infinitesimally different initial conditions may be completely uncorrelated with each other. Changes in the external forcing or variations in the boundary conditions can produce flows that vary from smooth laminar flow to more complicated motions with an identifiable length or time scale, and from there to the most complicated flow of them all, namely, fully developed turbulence with its spectrum of motions over many length scales. Depending on the specific system (for example, flow in a pipe or behind a grid), the transition from smooth laminar flow to fully developed turbulence may occur abruptly, or by successively more complex states, as the forcing is increased.

The difficulties of finding solutions to the Navier-Stokes equations that accurately predict and/or describe the transition to turbulence and the turbulent state itself are legendary, prompting the British physicist Sir Horace Lamb to remark, "I am an old man

[1] Reynolds rewrote the Navier-Stokes fluid equation as two equations—one for the mean velocity, which includes a quadratic term in the fluctuating velocity called the Reynolds stress, and one for the fluctuations, which is usually modeled by some suitable approximation. This approach underpins commonly used engineering models of turbulent fluid motion known as Reynolds-Averaged Navier-Stokes (RANS)—refer to Taylor (1938).

[2] The acceleration term looks complicated because of the advection term $\mathbf{u} \cdot \nabla \mathbf{u}$, which arises from the coordinate transformation from a frame moving with the fluid parcels (the "Lagrangian" frame, in which Newton's law has the usual form) to a frame of reference fixed in space (the "Eulerian" frame, in which other aspects of the mathematics are simpler). Specifically, acceleration of the fluid is, by definition, the second time derivative of the Lagrangian fluid trajectory $\mathbf{x}(t)$, which describes the motion of the fluid element that was initially at position $\mathbf{x}(0)$. The first time derivative is the Lagrangian fluid velocity, $d\mathbf{x}(t)/dt$, which is related to the Eulerian fluid velocity by $d\mathbf{x}(t)/dt = \mathbf{u}(t,\mathbf{x}(t))$. Because $\mathbf{u}$ is a function of time t and position $\mathbf{x}(t)$, which itself is a function of time, the Eulerian expression for the Lagrangian second time derivative (the fluid acceleration) is obtained through the chain rule and equals $d\mathbf{u}/dt = \partial\mathbf{u}/\partial t + \mathbf{u} \cdot \nabla \mathbf{u}$.

[3] Often, there is an additional term added to the right side of the equation that represents an external forcing of the flow per unit volume such as by gravity. Alternatively, the forcing can arise from the imposition of boundary conditions, whereby energy is injected by stresses at those boundaries.

Figure 1. Common Examples of Fluid Turbulence
Turbulence is commonly apparent in everyday life, as revealed by the collage of pictures above: (a) water flow from a faucet, (b) water from a garden hose, (c) flow past a curved wall, and (d) and (e) whitewater rapids whose turbulent fluctuations are so intense that air is entrained by the flow and produces small bubbles that diffusely reflect light and cause the water to appear white.

now, and when I die and go to heaven there are two matters on which I hope for enlightenment. One is quantum electrodynamics, and the other is the turbulent motion of fluids. And about the former I am rather optimistic"—1932 (in Tannehill et al. 1984). One of the most influential turbulence theorists in the last 40 years, Robert Kraichnan, started studying turbulence while working with Albert Einstein at Princeton, when he noticed the similarity between problems in gravitational field theory and classical fluid dynamics. His contributions include field-theoretic approaches to turbulence that have had recent stunning success when applied to the turbulent transport of passive scalar concentration (see the article "Field Theory and Statistical Hydrodynamics" on page 181).

What Is Turbulence?

So, what is turbulence and why is it so difficult to describe theoretically? In this article, we shall ignore the transition to turbulence and focus instead on fully developed turbulence. One of the most challenging aspects is that, in fully developed turbulence, the velocity fluctuates over a large range of coupled spatial and temporal scales. Examples of turbulence (Figure 1) are everywhere: the flow of water from a common faucet, water from a garden hose, the flow past a curved wall, and noisy rapids resulting from flow past rocks in an energetically flowing river. Another example is the dramatic pyroclastic flow in a volcanic eruption. In Figure 2, the explosive eruption of Mount St. Helens is illustrated at successively higher magnification, showing structure at many length scales. In all these examples, large velocity differences (as opposed to large velocities) resulting from shear forces applied to the fluid (or from intrinsic fluid instability) produce strong fluid turbulence, a state that can be defined as a solution of the Navier-Stokes equations whose statistics exhibit spatial and temporal fluctuations.

Historically, investigations of turbulence have progressed through alternating advances in experimental measurements, theoretical descriptions, and most recently, the introduction of numerical simulation of turbulence on high-speed computers. Similarly, there has been a rich interplay between fundamental understand-

ing and applications. For example, turbulence researchers in the early to mid 20th century were motivated by two important practical problems: predicting the weather and building ever more sophisticated aircraft. Aircraft development led to the construction of large wind tunnels, where measurements of the drag and lift on scaled model aircraft were used in the design of airplanes. On the other hand, weather prediction was severely hampered by the difficulty in doing numerical computation and was only made practical after the development, many decades later, of digital computers; in the early days, the calculation of the weather change in a day required weeks of calculation by hand! In addition to these two large problems, many other aspects of turbulent flow were investigated and attempts were made to factor in the effects of turbulence on the design and operation of real machines and devices.

To understand what turbulence is and why it makes a big difference in practical situations, we consider flow through a long cylindrical pipe of diameter D, a problem considered over a century ago by Osborne Reynolds (1894). Reynolds measured mean quantities such as the average flow rate and the mean pressure drop in the pipe. A practical concern is to determine the flow rate, or mean velocity U, as a function of the applied pressure, and its profile, as a function of distance from the wall. Because the fluid is incompressible, the volume of fluid entering any cross section of the pipe is the same as the volume flowing out of the pipe. Thus, the volume flux is constant along the flow direction. We can use Equation (1) to get a naive estimate of the mean velocity U for flow in a horizontal pipe. Consider as a concrete example the flow of water in a rigid pipe hooked up to the backyard water faucet. Taking a 3-meter length of a 2.5-centimeter diameter pipe and estimating the water pressure at 30 pounds per square inch (psi), the imposed pressure gradient ∇P is 0.1 psi/cm or 7000 dynes/cm^3. We assume the simplest case, namely, that the flow is smooth, or "laminar," so that the nonlinear term in Equation (1) can be neglected, and that the flow has reached its limiting velocity with $\partial U/\partial t = 0$. In that case, the density-normalized pressure gradient $\nabla P/\rho$ would be balanced by the viscous acceleration (or drag), $\nu\nabla^2\mathbf{u}$. Using dimensional arguments and taking into account that $\mathbf{u} = 0$ at the pipe wall, we estimate that $\nu\nabla^2\mathbf{u} \approx \nu U/D^2$, which yields the estimate for the mean flow velocity of $U \sim \nabla P D^2/\rho\nu$. Thus, for water with viscosity $\nu = 0.01$ cm^2/s flowing in a pipe with diameter $D = 2.5$ centimeters, the laminar flow velocity would reach $U \sim 40{,}000$ m/s or almost 30 times the speed of sound in water! Clearly, something is wrong with this argument. It turns out that the flow in such a pipe is turbulent (it has highly irregular spatial and temporal velocity fluctuations) and the measured mean flow velocity U is smaller by a factor of about 4000, or only about 10 m/s!

How can we improve our estimate? For the turbulent case, we might assume, as Reynolds did, that the nonlinear term dominates over the viscous term and then equate the nonlinear term ($\mathbf{u} \cdot \nabla\mathbf{u} \sim U^2/D$) to the pressure gradient, thereby obtaining the much more realistic estimate of $U \sim (\nabla P D/\rho)^{1/2} \sim 1.5$ m/s. This estimate actually overestimates the effects of the nonlinear term.

As illustrated in Figure 3, the solution for the laminar-flow velocity profile is quite gradual, whereas the turbulent velocity profile is much steeper at the walls and levels off in the center of the pipe. Evidently, the effect of turbulence is to greatly increase the momentum exchange in the central regions of the pipe, as large-scale eddies effectively 'lock up' the flow and thereby shift the velocity gradient (or velocity shear) closer to the wall. Because the flow resistance in the pipe is proportional to the steepness of the velocity profile near the wall, the practical consequence of the turbulence is a large increase in the flow resistance of the pipe—that is, less flow for a given applied pressure. The real-world implications of this increase in flow resistance are enormous: A large fraction of the world's energy consumption is devoted to compensating for turbulent energy loss! Nevertheless, the detailed understanding and prediction from first principles still elude turbulence theory.

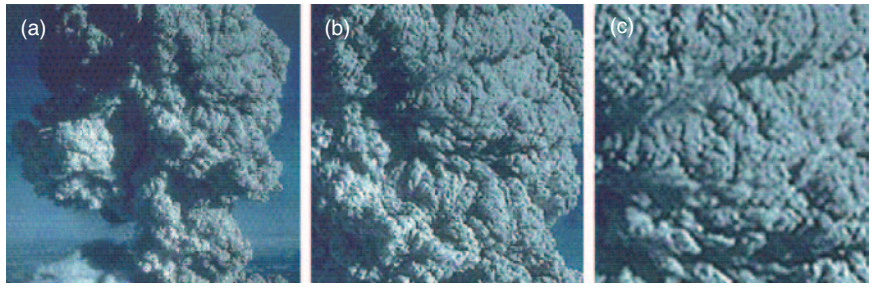


Figure 2. Scale-Independence in Turbulent Flows
The turbulent structure of the pyroclastic volcanic eruption of Mt. St. Helens shown in (a) is expanded by a factor of 2 in (b) and by another factor of 2 in (c). The characteristic scale of the plume is approximately 5 km. Note that the expanded images reveal the increasingly finer scale structure of the turbulent flow. The feature of scale independence, namely, that spatial images or temporal signals look the same (statistically) under increasing magnification is called self-similarity.

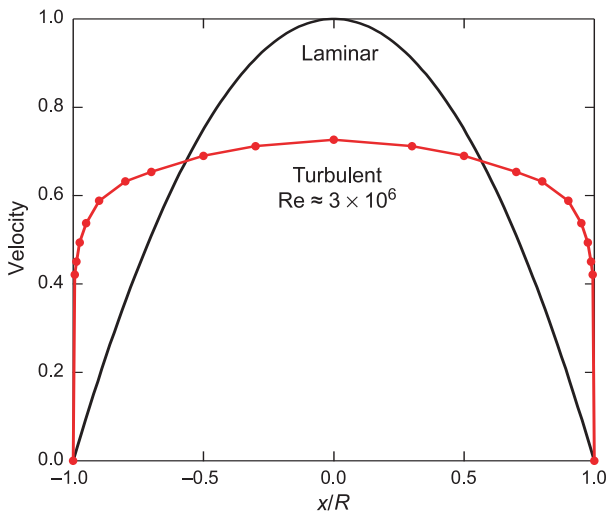


Figure 3. Mean Velocity Profiles for Laminar and Turbulent Pipe Flow
The velocity profile across the diameter ($D = 2R$ where R is the radius) of a pipe for laminar-flow conditions (black curve) shows a gradual velocity gradient compared with the very steep gradients near the walls resulting from turbulent flow conditions (red curve). Those steep gradients are proportional to the flow resistance in the pipe. Thus turbulence results in significantly less flow for a given applied pressure.

The example of pipe flow illustrates an important feature of turbulence—the ratio of the nonlinear term to the viscous dissipation term provides a good measure of the strength of turbulence. In fact, this ratio, $Re = UD/\nu$, where D is the size of the large-scale forcing (typically shear), is known as the Reynolds number after Reynolds' seminal work on pipe flow (1894). For a small Reynolds number, $Re << 1$, the nonlinearity can be neglected, and analytic solutions to the Navier-Stokes equation corresponding to laminar flow can often be found. When $Re >> 1$, however, there are no stable stationary solutions,[4] and the fluid flow is highly fluctuating in space and time, corresponding to turbulent flow. In particular, the flow is fully developed turbulence when Re is large compared with the Re for transition to turbulence for a particular set of forcing and boundary conditions. For example, in the problem above, where $D = 2.5$ cm and $U = 10$ m/s, the Reynolds number is $Re \sim 3 \times 10^5$ compared with a typical $Re \sim 2000$ for the onset of turbulence in pipe flow.

The Search for Universal Properties and the Kolmogorov Scaling Laws

In early laboratory experiments on turbulence, Reynolds and others supplemented their measurements of applied pressure and average velocity by observing the rapidly fluctuating character of the flow when they used dyes and other qualitative flow-visualization tools. In the atmosphere, however, one could measure much longer-term fluctuations, at a fixed location, and such Eulerian measurements intrigued the young theoretical physicist G. I. Taylor (1938). Turbulence is difficult to measure because the turbulent state changes rapidly in space and time. Taylor proposed a probabilistic/statistical approach based on averaging over ensembles of individual realizations, although he soon replaced ensemble averages by time averages at a fixed point in space. Taylor also used the idealized concept (originally introduced by Lord Kelvin in 1887) of statistically homogeneous, isotropic turbulence. Homogeneity and isotropy imply that spatial translations and rotations, respectively, do not change the average values of physical variables.[5]

Lewis F. Richardson was another influential fluid dynamicist of the early 20th century. Richardson performed the first numerical computation for predicting the weather (on a hand calculator)! He also proposed (1926) a pictorial description of turbulence called a cascade, in which nonlinearity transforms large-scale velocity circulations (or eddies, or whorls) into circulations at successively smaller scales (sizes) until they reach such a small scale that the circulation of the eddies is efficiently dissipated into heat by viscosity. Richardson captured this energy cascade in a poetic take-off on Jonathan Swift's famous description of fleas.[6] In Richardson's words, "Big whorls have little whorls that feed on their velocity, and little whorls have lesser whorls and so on to viscosity" (circa 1922). A schematic illustration of the energy cascade picture is shown in Figure 4, where the mean energy

[4] How large Re must be to get a turbulent state depends on the particular source of forcing and on the boundary conditions. For example, the transition to turbulence in pipe flow can occur anywhere in the range $1000 < Re < 50{,}000$ depending on inlet boundary conditions and the roughness of the pipe wall. For most commonly encountered conditions, the transition is near $Re = 2000$.

[5] Many theoretical descriptions use these assumptions, but typical turbulence encountered in the real world often obeys neither condition at large scales. A key question in real-world situations is whether the assumptions of homogeneity and isotropy are satisfied at small scales, thus justifying application of a general framework for those smaller scales.

[6] "So, the naturalists observe, the flea,/ hath smaller fleas that on him prey;/ And these have smaller still to bite 'em;/ And so proceed, ad infinitum."—Jonathan Swift, *Poetry, a Rhapsody*

injection rate ε at large scales is balanced by the mean energy dissipation rate at small scales. Richardson and Taylor also appreciated that generic properties of turbulence may be discovered in the statistics of velocity differences between two locations separated by a distance **r**, denoted as $\delta\mathbf{u}(\mathbf{x}, \mathbf{x}+\mathbf{r}) = \mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x}+\mathbf{r})$. The statistics of velocity differences at two locations are an improvement over the statistics of velocity fluctuations at a single location for a number of technical reasons, which we do not discuss here. Longitudinal projections of velocity differences

$$\delta u(r) = \delta\mathbf{u}(\mathbf{x}, \mathbf{x}+\mathbf{r}) \cdot \mathbf{r} / |\mathbf{r}|$$

are often measured in modern experiments and are one of the main quantities of interest in the analysis of fluid turbulence.

Measuring velocity differences on fast time scales and with high precision was a difficult proposition in the early 20th century and required the development of the hot-wire anemometer, in which fluid flowing past a thin heated wire carries heat away at a rate, proportional to the fluid velocity. Hot-wire anemometry made possible the measurement, on laboratory time scales, of the fluctuating turbulent velocity field at a single point in space (see Figure 5). For a flow whose mean velocity is large, velocity differences were inferred from those single-point measurements by Taylor's "frozen-turbulence" hypothesis.[7]

[7] If the mean velocity is large compared with velocity fluctuations, the turbulence can be considered "frozen" in the sense that velocity fluctuations are swept past a single point faster than they would change because of turbulent dynamics. In that case, the spatial separation Δr is related to the time increment Δt by $\Delta r = -U\Delta t$, where U is the mean velocity. See also the article "Taylor's Hypothesis, Hamilton's Principle, and the LANS-α Model for Computing Turbulence" on page 152 .

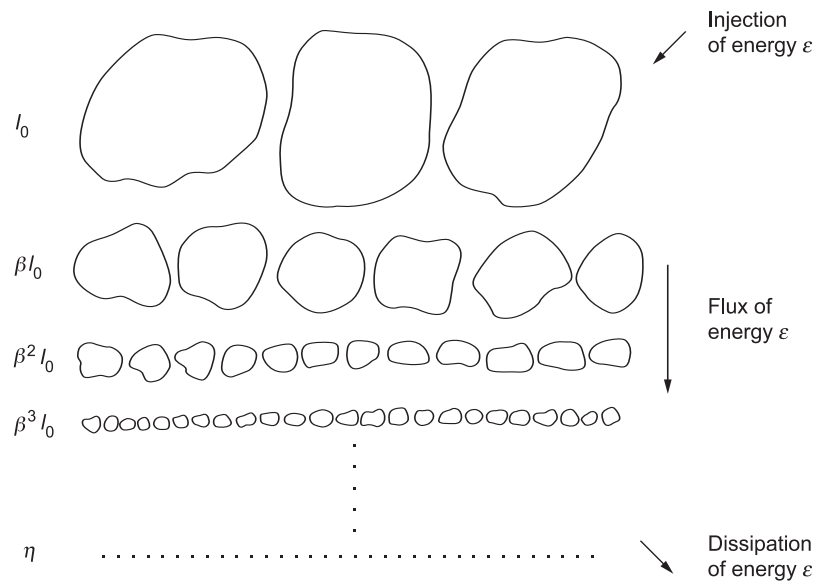


Figure 4. The Energy Cascade Picture of Turbulence
This figure represents a one-dimensional simplification of the cascade process with β representing the scale factor (usually taken to be 1/2 because of the quadratic nonlinearity in the Navier-Stokes equation). The eddies are purposely shown to be "space filling" in a lateral sense as they decrease in size.
(This figure was modified with permission from Uriel Frisch. 1995. *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge, UK: Cambridge University Press.)

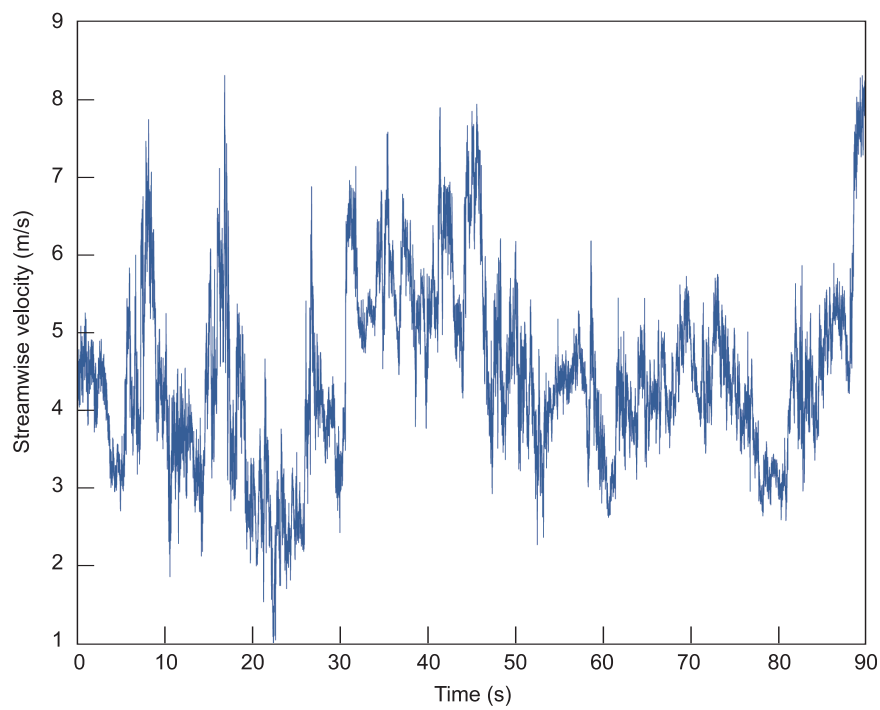


Figure 5. Time Series of Velocities in a Turbulent Boundary Layer
This time series of velocities for an atmospheric turbulent boundary layer with Reynolds number $Re \sim 2 \times 10^7$ was measured at a single location with a hot-wire anemometer. The velocity fluctuations are apparently random. (This figure is courtesy of Susan Kurien of Los Alamos, who has used data recorded in 1997 by Brindesh Dhruva of Yale University.)

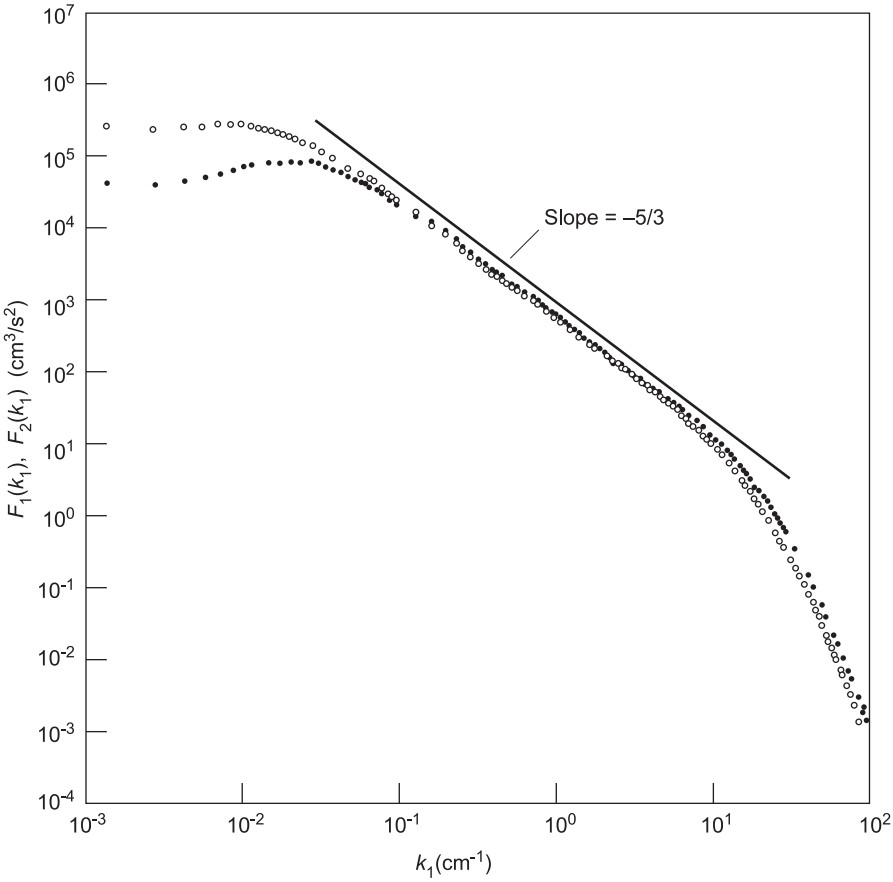


Figure 6. Kolmogorov-like Energy Spectrum for a Turbulent Jet
The graph shows experimental data for the energy spectrum (computed from velocity times series like that in Figure 5) as a function of wave number *k*, or *E*(*k*) vs *k*, for a turbulent jet with Reynolds number *Re* ~ 24,000. Note that the measured spectrum goes as $E(k) \propto k^{-5/3}$. (Champagne 1978. Redrawn with the permission of Cambridge University Press.)

Single-point measurements of turbulent velocity fluctuations have been performed for many systems and have contributed both to a fundamental understanding of turbulence and to engineering modeling of the effects of turbulence at small scales on the flow at larger scales. (See the section on engineering models.) Single-point measurements of velocity fluctuations have been the primary tool for investigating fluid turbulence.They remain in common use because of their large dynamic range and high signal-to-noise ratio relative to more modern developments such as particle image velocimetry, in which the goal is to measure whole velocity fields. For now, we consider results that were motivated or measured with the limitations of single-point experiments in mind.

The Kolmogorov Scaling Laws. In 1938, von Kármán and Howarth derived an exact theoretical relation for the dynamics of turbulence statistics. Starting from the Navier-Stokes equation and assuming homogeneity and isotropy, the two scientists derived an equation for the dynamics of the lowest-order two-point velocity correlation function. (This function is $\langle \mathbf{u}(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x}+\mathbf{r}) \rangle$, where the angle brackets denote an ensemble average, that is, an average over many statistically independent realizations of the flow. The two-point velocity correlation functions cannot describe universal features of turbulence because they are scale dependent (the large-scale flow dominates their behavior) and they lack Galilean invariance. Nevertheless, their derivation inspired a real breakthrough. In 1941, Andrei Kolmogorov recast the Kármán-Howarth equation in terms of the moments of $\delta u(r)$, the velocity differences across scales, thereby producing a relationship between the second moment $\langle [\delta u(r)]^2 \rangle$ and the third moment $\langle [\delta u(r)]^3 \rangle$. These statistical objects, which retain Galilean invariance and hence hold the promise of universality, are now known as structure functions.

Kolmogorov then proposed the notion of an "inertial range" of scales based on Richardson's picture of the energy cascade: Kinetic energy is injected at the largest scales L of the flow at an average rate ε and generates large-scale fluctuations. The injected energy cascades to smaller scales via nonlinear inertial (energy-conserving) processes until it reaches a scale of order ℓ_d, where viscous dissipation becomes dominant and the kinetic energy is converted into heat. In other words, the intermediate spatial scales r, in the interval $\ell_d << r << L$, define an inertial range in which large-scale forcing and viscous forces have negligible effects. With these assumptions and the Kármán-Howarth equation recast for structure functions, Kolmogorov derived the famous "four-fifths law." The equation defining this law describes an exact relationship for the third-order structure function within the inertial range:

$$\left\langle \left[\delta u(r)\right]^3 \right\rangle = -(4/5)\varepsilon r \ ,$$

where ε is assumed to be the finite energy-dissipation rate (per unit mass) of the turbulent state. This relation-

ship is a statement of conservation of energy in the inertial range of scales of a turbulent fluid; the third moment, which arises from the nonlinear term in Equation (1), is thus an indirect measure of the flux of energy through spatial scales of size r. (High Reynold's-number numerical simulations are compared with the four-fifths law and with other statistical characterizations of turbulence in the article "Direct Numerical Simulations of Turbulence" on page 142.)

Kolmogorov further assumed that the cascade process occurs in a self-similar way. That is, eddies of a given size behave statistically the same as eddies of a different size. This assumption, along with the four-fifths law, gave rise to the general scaling prediction of Kolmogorov, which states that the nth order structure function (referred to in the article "Direct Numerical Simulations of Turbulence" as $S_n(r)$) must scale as $r^{n/3}$. During the decades that have passed since Kolmogorov's seminal papers (1941), empirical departures from his scaling prediction have been measured for n different from 3, leading to our present understanding that turbulent scales are not self-similar, but that they become increasingly intermittent as the scale size decreases. The characterization and understanding of these deviations, known as the "anomalous" scaling feature of turbulence, have been of sustained and current interest (see the box "Intermittency and Anomalous Scaling in Turbulence" on page 136).

Empirical observations show that a flow becomes fully turbulent only when a large range of scales separates the injection scale L and the dissipation scale ℓ_d. A convenient measure of this range of spatial scales for fluid turbulence, which also characterizes the number of degrees of freedom of the turbulent state, is the large-scale Reynolds number, Re. The Reynolds number is also the ratio of nonlinear to viscous forces introduced earlier in the context of pipe flow. Most theories of turbulence deal with asymptotically large Re, that is, $Re \to \infty$, so that an arbitrarily large range of scales separates the injection scales from the dissipation scales.

Because energy cascading down through spatial scales is a central feature of fluid turbulence, it is natural to consider the distribution of energy among spatial scales in wave number (or Fourier) space, as suggested by Taylor (1938). The energy distribution $E(\boldsymbol{k}) = 1/2|\tilde{\mathbf{u}}(\mathbf{k})|^2$, where $\tilde{\mathbf{u}}(\mathbf{k})$ is the Fourier transform of the velocity field and the wave number k is related to the spatial scale ℓ by $k = 2\pi/\ell$.[8] Wave-number space is very useful for the representation of fluid turbulence because differential operators in real space transform to multiplicative operators in k-space. For example, the diffusion operator in the term $\nu\nabla^2\mathbf{u}$ becomes $\nu k^2\tilde{\mathbf{u}}$ in the Fourier representation. Another appealing feature of the wave number representation is the nonlocal property of the Fourier transform, which causes each Fourier mode represented by wave number k to represent cleanly the corresponding scale ℓ. On the other hand, the k-space representation is difficult from the perspective of understanding how spatial structures, such as intense eddies, affect the transfer of energy between scales, that is between eddies of different sizes.

The consequence of energy conservation on the form of $E(k)$ was independently discovered by Obukov (1941), Heisenberg (1948), and Onsager (1949), all of whom obtained the scaling relationship for the energy spectrum $E(k) \sim k^{-5/3}$ for the inertial scales in fully developed homogeneous isotropic turbulence. This result is not independent of the picture presented above in terms of real space-velocity differences but is another way of looking at the consequences of energy conservation. Many subsequent experiments and numerical simulations have observed this relationship to within experimental/numerical uncertainty, thereby lending credence to the energy cascade picture. Figure 6 shows the energy spectrum obtained from time series measurements at a single point in a turbulent jet, where the spatial scale is related to time by Taylor's hypothesis that the large mean velocity sweeps the "frozen-in" turbulent field past the measurement point.

Vorticity. Another important quantity in the characterization and understanding of fluid turbulence is the vorticity field, $\boldsymbol{\omega}(x,t) = \nabla \times \mathbf{u}(x,t)$, which roughly measures the local swirl of the flow as picturesquely drawn by da Vinci in the opening illustration. The notion of an "eddy" or "whorl" is naturally associated with one's idea of a vortex—a compact swirling object such as a dust devil, a tornado, or a hurricane—but this association is schematic at best. In three-dimensional (3-D) turbulence, vorticity plays a quantitative role in that the average rate of energy dissipation ε is related to the mean-square vorticity by the relation $\varepsilon = -u\langle |\boldsymbol{\omega}|^2\rangle$. Vorticity plays a different role in two-dimensional (2-D) turbulence. Vortex stretching has long been recognized as an important ingredient in fluid turbulence (Taylor 1938); if a vortex tube is stretched so that its cross section is reduced, the mean-square vorticity in that cross section will increase, thereby causing strong spatially localized vortex filaments that dissipate energy. The notion of vortex stretching and energy dissipation is discussed in the article "The LANS-α model for Computing Turbulence" on page 152.

[8]If isotropy is assumed, the energy distribution $E(\mathbf{k})$, where $\mathbf{k}$ is a vector quantity, depends only on the magnitude $k = |\mathbf{k}|$ and one can denote the energy as $E(k)$ without loss of generality.

Engineering Models of Turbulence

It is worth stressing again that turbulence is both fundamentally interesting and of tremendous practical importance. As mentioned above, modeling complex practical problems requires a perspective different from that needed for studying fundamental issues. Foremost is the ability to get fairly accurate results with minimal computational effort. That goal can often be accomplished by making simple models for the effects of turbulence and adjusting coefficients in the model by fitting computational results to experimental data. Provided that the parameter range used in the model is well covered by experimental data, this approach is very efficient. Examples that have benefited from copious amounts of reliable data are aircraft design—aerodynamics of body and wing design have been at the heart of a huge international industry—and aerodynamic drag reduction for automobiles to achieve better fuel efficiency. Global climate modeling and the design of nuclear weapons, on the other hand, are examples for which data are either impossible or quite difficult to obtain. In such situations, the utmost care needs to be taken when one attempts to extrapolate models to circumstances outside the validation regime.

The main goal of many engineering models is to estimate transport properties—not just the net transport of energy and momentum by a single fluid but the transport of matter such as pollutants in the atmosphere or the mixing of one material with another. Mixing is a critical process in inertial confinement fusion and in weapons physics applications. It is crucial for certification of the nuclear weapons stockpile that scientists know how well engineering models are performing and use that knowledge to predict outcomes with a known degree of certainty.

The Closure Problem for Engineering Models. Engineering models are constructed for computational efficiency rather than perfect representation of turbulence. The class of engineering models known as Reynolds-Averaged Navier-Stokes (RANS) provides a good example of how the problem of "closure" arises and the parameters that need to be determined experimentally to make those models work. Consider again the flow of a fluid with viscosity ν in a pipe with a pressure gradient along the pipe. When the pressure applied at the pipe inlet and the pipe diameter D are small, the fluid flow is laminar, and the velocity profile in the pipe is quadratic with a peak velocity U that is proportional to the applied pressure (refer to Figure 3). When the forcing pressure gets large enough to produce a high flow velocity and therefore a large Reynolds number, typically $Re = UD/\nu \geq 2000$, the flow becomes turbulent, large velocity fluctuations are present in the flow, and the velocity profile changes substantially (refer again to Figure 3). An engineering challenge is to compute the spatial distribution of the mean velocity of the turbulent flow. Following the procedure first written down by Reynolds, the velocity and pressure fields are separated into mean (the overbar denotes a time average) and fluctuating (denoted by the prime) parts:

$$u_i(\mathbf{x},t) = \bar{u}_i(\mathbf{x},t) + u_i'(\mathbf{x},t) \ ,$$

where i denotes one of the components of the vector field $\mathbf{u}(x,t)$ and the average of the fluctuating part of the velocity is zero by definition. Also,

$$P(\mathbf{x},t) = \bar{P}(\mathbf{x},t) + P'(\mathbf{x},t) \ .$$

Substituting these expressions into Equation (1), using the constant-density continuity condition

$$\nabla \cdot \mathbf{u} = 0 \Rightarrow \nabla \cdot \bar{\mathbf{u}} = 0 = \nabla \cdot \mathbf{u}' \ ,$$

and averaging term by term yields an equation for the mean velocity:

$$\frac{\partial \bar{u}_i}{\partial t} + \bar{u}_j \frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \overline{u_i' u_j'}}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \bar{P}}{\partial x_i} + \nu \frac{\partial^2 \bar{u}_i}{\partial x_j^2} \ . \tag{2}$$

Note that the equation for the mean flow looks the same as the Navier-Stokes equation for the full velocity $\mathbf{u}$, Equation (1), with the addition of a term involving the time average of a product of the fluctuating parts of the velocity, namely, the Reynolds stress tensor,

$$R_{ij} = \overline{u_i' u_j'} \ ,$$

That additional term, which represents the transport of momentum caused by turbulent fluctuations, acts like an effective stress on the flow and cannot, at this time, be determined completely from first principles. As a result, many schemes have been developed to approximate the Reynolds stress.

The simplest formulation for the Reynolds stress tensor is

$$R_{ij} = -\nu_{\mathrm{T}}(x) \frac{\partial \bar{u}_i}{\partial x_j} \ ,$$

where $\nu_{\mathrm{T}}(x)$ is called the turbulent eddy viscosity because the additional term looks like a viscous diffusion term. A more sophisticated approach is to solve for the Reynolds stress by writing an equation for the time evolution of $\overline{u_i' u_j'}$ (Johnson et al. 1986). This equation has multiple undetermined coefficients and depends on the third moment $\overline{u_i' u_j' u_k'}$. Again, the third-order moment is unknown and needs to be approximated or written in terms of fourth-order moments. In principle, an infinite set of equations for higher-order moments is required, so one needs to "close" the set at a small number to achieve computational effi-

ciency. At any stage of approximation, undetermined coefficients are set by comparison with experimental or direct numerical simulation data. This approach is often very effective, although it does depend on the quality of the data and on the operating parameter regime covered by the data.

Modern Developments

By the end of the 1940s, great progress had been made in the study of turbulence. The statistical approach to turbulence, the importance of the energy and its wave number representation, the notion of measuring velocity differences, and the dynamics of vortex structures as an explanation of the mechanism of fluid turbulence had all been articulated by Taylor. The cascade picture of Richardson had been made quantitative by Kolmogorov and others. The concepts of universal scaling and self-similarity were key ingredients in that description. On the practical side, tremendous advances had been made in aeronautics, with newly emerging jet aircraft flying at speeds approaching the speed of sound. Empirical models based on the engineering approach described above were being used to describe these practical turbulence problems.

The next 50 years were marked by steady progress in theory and modeling, increasingly sophisticated experiments, and the introduction and widespread use of the digital computer as a tool for the study of turbulence. In the remainder of this review, we touch on some of the advances of the post-Kolmogorov era, paying particular attention to the ever-increasing impact of the digital computer on three aspects of turbulence research: direct numerical simulations of idealized turbulence, increasingly sophisticated engineering models of turbulence, and the extraordinary enhancement in the quality and quantity of experimental data achieved by computer data acquisition. As far back as the Manhattan Project, the computer (more exactly, numerical schemes implemented on a roomful of Marchand calculators) began to play a major role in the calculations of fluid problems. A leading figure in that project, John von Neumann (1963), noted in a 1949 review of turbulence that "… a considerable mathematical effort towards a detailed understanding of the mechanism of turbulence is called for" but that, given the analytic difficulties presented by the turbulence problem, "… there might be some hope to 'break the deadlock' by extensive, but well-planned, computational efforts." Von Neumann's foresight in understanding the important role of computers for the study of turbulence predated the first direct numerical simulation of the turbulent state by more than 20 years.

From a fundamental perspective, the direct numerical simulation of idealized isotropic, homogeneous turbulence has been revolutionary in its impact on turbulence research because of the ability to simulate and display the full 3-D velocity field at increasingly large Reynolds number. Similarly, experimentation on turbulence has advanced tremendously by using computer data acquisition; 20 years ago it was possible to measure and analyze time series data from single-point probes that totaled no more than 10 megabytes of information, whereas today statistical ensembles of thousands of spatially and temporally resolved velocity fields, taking 10 terabytes of storage space can be obtained and processed. This millionfold increase in experimental capability has opened the door to great new possibilities in turbulence research that will be enhanced even further by expected future increases in computational power.

Below, we briefly address advances in numerical simulation, in turbulence modeling, and theoretical understanding of passive scalar transport, topics dealt with more extensively in the articles immediately following this one. We then describe several exciting new experimental advances in fluid turbulence research and close this introduction with a view toward "solving" at least some aspect of the turbulence problem.

Direct Numerical Simulation of Turbulence

Recent advances in large-scale scientific computing have made possible direct numerical simulations of the Navier-Stokes equation under turbulent conditions. In other words, for simulations performed on the world's largest supercomputers, no closure or subgrid approximations are used to simplify the flow, but rather the simulated flow follows all the twisting-turning and stretching-folding motions of the full-blown Navier-Stokes equations at effective large-scale Reynolds numbers of about 10^5. These simulations render available for analysis the entire 3-D velocity field down to the dissipation scale. With these numerically generated data, one can study the structures of the flow and correlate them with turbulent transfer processes, the nonlinear processes that carry energy from large to small scales.

An especially efficient technique for studying isotropic, homogeneous turbulence is to consider the flow in a box of length L with periodic boundary conditions and use the spectral method, an orthogonal decomposition into Fourier modes, to simulate the Navier-Stokes equation. Forcing is typically generated by maintaining constant energy in the smallest k mode (or a small number of the longest-wavelength modes).

The first direct numerical simulation of fluid turbulence (Orszag and Patterson 1972) had a resolution of 32^3, corresponding to $Re \sim 100$. By the early 1990s, Reynolds numbers of about 6000 for a 512^3 simulation could be obtained (She et al. 1993); the separation between the box size and the dissipation scale was just short of a decade. Recent calculations on the Los Alamos Q machine, using 2048^3 spatial resolution, and on the Japanese Earth Simulator with 4096^3 modes (Kaneda et al. 2003) achieved a Reynolds number of about 10^5, corresponding to about 1.5 decades of turbulent scales, which is approaching fully developed turbulence in modestly sized wind tunnels. It is important to appreciate that the *Re* of direct numerical turbulence simulations grows only as $Re \propto N^{4/9}$, where N is the number of degrees of freedom that are computed: A factor of 2 increase in the linear dimension of the box means computing 2^3 more modes for a corresponding increase in *Re* of about 2.5. Nevertheless, for isotropic, homogeneous fully developed turbulence, direct numerical simulation has become the tool of choice for detailed characterization of fundamental flow properties and comparison with Kolmogorov-type theories. More details regarding numerical simulation of turbulence can be found in the article "Direct Numerical Simulations of Turbulence" on page 142.

Modern Turbulence Models

Although the RANS models described above maintain a dominant role in turbulence modeling, other approaches have become tractable because of increases in computational power. A more recent approach to modeling turbulent processes is to decompose spatial scales of the flow into Fourier modes and then to truncate the expansion at some intermediate scale (usually with a smooth Gaussian filter) and model the small scales with a subgrid model. One then computes the large scales explicitly and approximates the effect of the small scales with the subgrid model. This class of methods (Meneveau and Katz 2000) is called large eddy simulation (LES) and has become an alternative to RANS modeling when more-accurate spatial information is required. Because of the spatial filtering, LES modeling has problems with boundaries and is less computationally efficient than RANS techniques. Nevertheless, LES models may be more universal than RANS models and therefore rely less on ad hoc coefficients determined from experimental data.

Another variant of the subgrid-model approach recently invented at Los Alamos is the Lagrangian-averaged Navier-Stokes alpha (LANS-α) model. Although not obtainable by filtering the Navier-Stokes equations, the LANS-α model has a spatial cutoff determined by the coefficient α. For spatial scales larger than α, the dynamics are computed exactly (in effect, the Navier-Stokes equations are used) and yield the energy spectrum $E(k) \propto k^{-5/3}$, whereas for spatial scales less than α, the energy falls more rapidly, $E(k) \propto k^{-3}$. The LANS-α model can be derived from a Lagrangian-averaging procedure starting from Hamilton's principle of least action. It is the first-closure (or subgrid) scheme to modify the nonlinear term rather than the dissipation term and, as a result, has some unique advantages relative to more traditional LES schemes (see the articles "The LANS-α Model for Computing Turbulence" and "Taylor's Hypothesis, Hamilton's Principle, and the LANS-α Model for Computing Turbulence" on pages 152 and 172, respectively).

Beyond the Kolmogorov Theory

The 50 years that have passed, from about 1950 until the new millennium, were notable for increasingly sophisticated theoretical descriptions of fluid turbulence, including the seminal contributions of Robert Kraichnan (1965, 1967, 1968, 1975), who pioneered the foundations of the modern statistical field-theory approach to hydrodynamics, particularly by predicting the inverse energy cascade from small scales to large scales in 2-D turbulence, and George Batchelor (1952, 1953, 1959, 1969). Those developments are generally beyond the scope of the present review, and we have already referred the reader to recent books in which they are surveyed in detail (McComb 1990, Frisch 1995, Lesieur 1997). We touch briefly, however, on a few recent developments that grew out of those efforts and on the influence of the Lagrangian perspective of fluid turbulence.

The physical picture that emerges from the Kolmogorov phenomenology is that the turbulent scales of motion are self-similar; that is, the statistical features of the system are independent of spatial scale. One measure of this self-similarity is the nondimensional ratio of the fourth moment to the square of the second moment, $F = \langle \delta u^4 \rangle / \langle \delta u^2 \rangle^2$, as a function of scale separation. If the velocity distribution is self-similar, then F should be constant, or flat, as a function of length scale. Indeed, the nondimensional ratio of any combination of velocity-increment moments should be scale independent. If, however, F behaves as a power law in the separation r, then the system is not self-similar, but instead it is characterized by intermittency: short bursts (in time) or isolated regions (in space) of high-amplitude fluctuations separated by relatively quiescent periods or

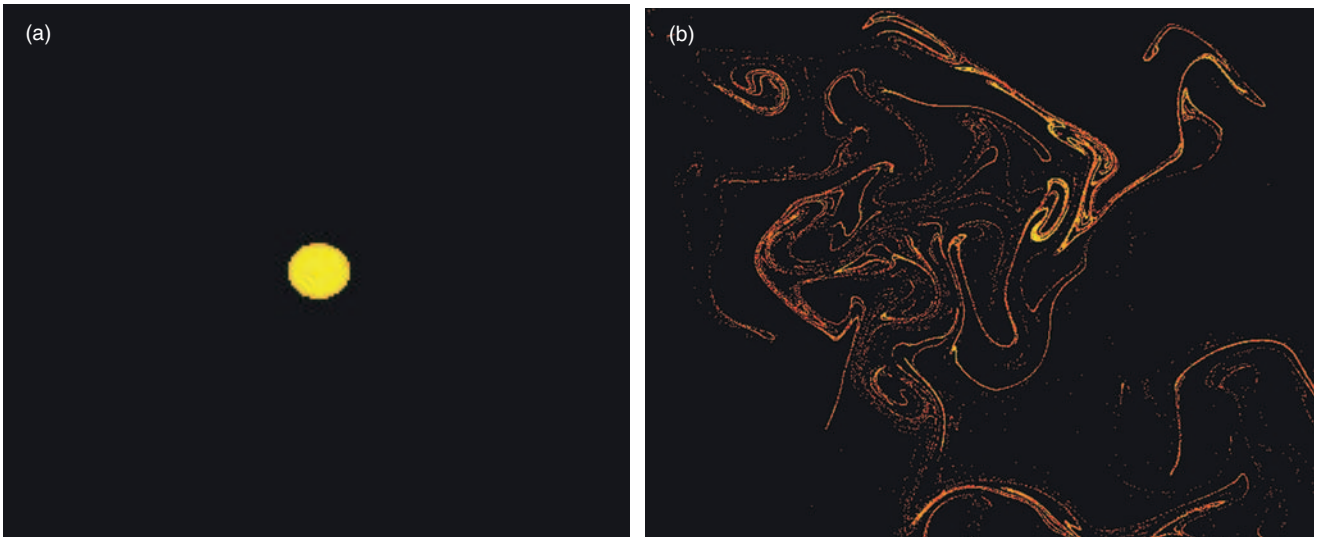


Figure 7. Passive Scalar Turbulence in a Stratified Layer
An effective blob of yellow dye is carried by a forced 2-D turbulent flow in a stratified layer. The images show (a) the initial dye concentration and (b) the concentration after about one rotation of a large-scale eddy. The sharp gradients in the concentration lead to the very strong anomalous scaling in the transport of the passive scalar field.

regions. From the 1960s to the 1980s, experimentalists reported departures from the Kolmogorov scaling. The measured fourth-order and higher moments of velocity differences did not scale as $r^{n/3}$, but rather as a lower power of the separation r, $\langle \delta u^n \rangle \sim r^{\xi_n}$, with $\xi_n < n/3$ for $n > 3$. To preserve the correct dimensions of the nth-order velocity difference moments, the deviations from Kolmogorov scaling, or from self-similarity, can be written as a correction factor given by the ratio of the large scale L to the separation r to some power Δ_n, or $(L/r)^{\Delta_n}$ (see the box "Intermittency and Anomalous Scaling in Turbulence" on page 136). Some recent analytic progress toward understanding the origin of the observed anomalous scaling has been made in the context of passive scalar turbulence and involves the application of nonperturbative field-theory techniques to that problem (see the article "Field Theory and Statistical Hydrodynamics" on page 181).

The passive scalar problem describes the transport and effective diffusion of material by a turbulent velocity field. This stirring process characterizes fluid mixing, which has many important scientific and technical applications. Whereas intermittency is rather weak in turbulent velocity statistics, the distribution of a passive scalar concentration carried by a turbulent flow is very intermittent. In other words, there is a much larger probability (compared with what one would expect for a random, or Gaussian, distribution) of finding local concentrations that differ greatly from the mean value. For characterizing fluid mixing, the Lagrangian frame of reference (which moves with the fluid element as opposed to the Eulerian frame, which is fixed in space) is very useful theoretically because a passive scalar is carried by fluid elements. Figure 7 shows the distribution of a virtual drop of yellow dye carried by a 2-D turbulent flow in a stratified layer experiment.[9] The structured distribution of the dye illustrates how the velocity field stretches and folds fluid elements to produce mixing. Adopting the Lagrangian frame of reference is rapidly emerging as a powerful new approach for modeling turbulent mechanisms of energy transfer. This approach has led to Lagrangian tetrad methods, a phenomenological model arising from the nonperturbative field-theoretical approach to turbulence, and to the LANS-α model mentioned above.

[9]The "virtual" drop consists of more than 10,000 fluid elements, whose evolution is computed by solving the Lagrangian equation $d\mathbf{x}(t)/dt = \mathbf{u}(t,\mathbf{x}(t))$ from experimental velocity fields.

Recent Experimental Developments

Quantitative single-point measurements of velocity combined with qualitative flow visualization (van Dyke 1982) have characterized almost all experimental measurements of fluid turbulence for most of the 20th century. Recently, however, new experimental techniques enabled by digital data acquisition, powerful pulsed-laser technology, and fast digital imaging systems have emerged and are causing great excitement in the field of turbulence.

Intermittency and Anomalous Scaling in Turbulence

Misha Chertkov

Intermittency is associated with the violent, atypical discontinuous nature of turbulence. When a signal from turbulent flow (for example, the velocity along a particular direction) is measured at a single spatial point and a sequence of times (an Eulerian measurement), the fluctuations in the values appear to be random. Since any random sequence is most naturally explained in terms of its statistical distribution, one typically determines the statistics by constructing a histogram of the signal. The violent nature of turbulence shows itself in the very special shape of the histogram, or the probability distribution function (PDF), of the turbulent velocity signal. It is typically wider than the Gaussian distributions emerging in the context of equilibrium statistical physics—for example, the Gaussian distribution that describes the velocity of (or the distance traveled by) a molecule undergoing Brownian motion.

The PDF of the energy dissipation rate, $P(\varepsilon)$, illustrates how far from Gaussian a turbulent distribution can be. At values far above the average, $\varepsilon >> \langle\varepsilon\rangle$, where $\varepsilon \equiv \nu(\nabla u)^2$, the probability distribution $P(\varepsilon)$ has a stretched exponential tail, $\ln P(\varepsilon) \propto -\varepsilon^a$ (La Porta 2001). The extended tail of the turbulent PDF illustrates the important role played by the atypical, violent, and rare events in turbulence.

Intermittency has many faces. In the context of two-point measurements, intermittency is associated with the notion of anomalous scaling. Statistics of the longitudinal velocity increments, $\delta u(r)$ (the difference in the velocity components parallel to the line separating the two points) in developed turbulence becomes extremely non-Gaussian as the scale decreases. In particular, if the scale r separating the two points is deep inside the inertial interval, $L >> r >> \ell_d$, then the nth moment of the longitudinal velocity increment is given by

$$\left\langle [\delta u(r)]^n \right\rangle \approx \langle\varepsilon\rangle^{n/3} r^{n\alpha/3} \left(L/r\right)^{\Delta_n} , \qquad (1)$$

where L is the integral (pumping, energy-containing) scale of turbulence. The first thing to mention about Equation (1) is that the viscous, Kolmogorov scale ℓ_d does not enter the relation in the developed turbulence regime. This fact is simply related to the direction of the energy cascade: On average, energy flows from the large scale, where it is pumped into the system, toward the smaller scale, ℓ_d, where it is dissipated; it does not flow from the small scale. Secondly, Δ_n on the right side of Equation (1) is the anomalous scaling exponent. In the phenomenology proposed by Kolmogorov in 1941, the flow is assumed to be self-similar in the inertial range of scales, which implies that anomalous scaling is absent, $\Delta_n = 0$, for all values of n. The self-similar scaling phenomenology is an extension of the four-fifths law proven by Kolmogorov in 1941 for the third moment

$$\left\langle [\delta u(r)]^3 \right\rangle = -\frac{4}{5}\langle\varepsilon\rangle r \ .$$

(See discussion of the four-fifths law in "Direct Numerical Simulations of Turbulence" on page 142). This law is a statement of conservation of energy from scale to scale in the inertial regime of homogeneous isotropic turbulence. Modern experimental and numerical tests (Frisch 1995) unequivocally dismiss the self-similarity assumption, $\Delta_n = 0$, as invalid. But so far, theory is still incapable of adding any other exact relation to the celebrated four-fifths law.

On the other hand, even though a comprehensive theoretical analysis of developed isotropic turbulence remains elusive, there has been an important breakthrough in understanding anomalous scaling in the simpler problem of passive scalar turbulence (see the article "Field Theory and Statistical Hydrodynamics" on page 181).

One innovative example of a new turbulence experiment (La Porta 2001) is the use of silicon strip detectors from high-energy physics to track a single small particle in a turbulent flow with high Reynolds number (see Figure 8). This is an example of the direct measurement of Lagrangian (moving with the fluid) properties of the fluid flow. Because the particle trajectories are time resolved, the acceleration statistics can be obtained directly from experiment, and theoretical predictions for those statistics can be tested.

Another application of new technology has made possible the local time-resolved determination of the full 3-D velocity gradient tensor[10] at a point in space (Zeff 2003). Knowing the local velocity gradients allows one to calculate the energy dissipation rate ε and mean-square vorticity, $\Omega = \langle \omega^2 \rangle/2$, and thereby provide an experimental measure of intense and intermittent dissipation events (see Figure 9).

Both these techniques can be used to obtain large data samples that are statistically converged and have on the order of 10^6 data sets per parameter value. At present, however, the physical length scales accessible to these two techniques are constrained to lie within or close to the dissipation scale ℓ_d. By using holographic methods, one can obtain highly resolved, fully 3-D velocity fields (see Figure 10), which allow the full turbulent inertial range of scales to be investigated (Zhang et al. 1997). For holographic measurements, however, one is limited to a small number of such realizations, and time-resolved measurements are not currently achievable.

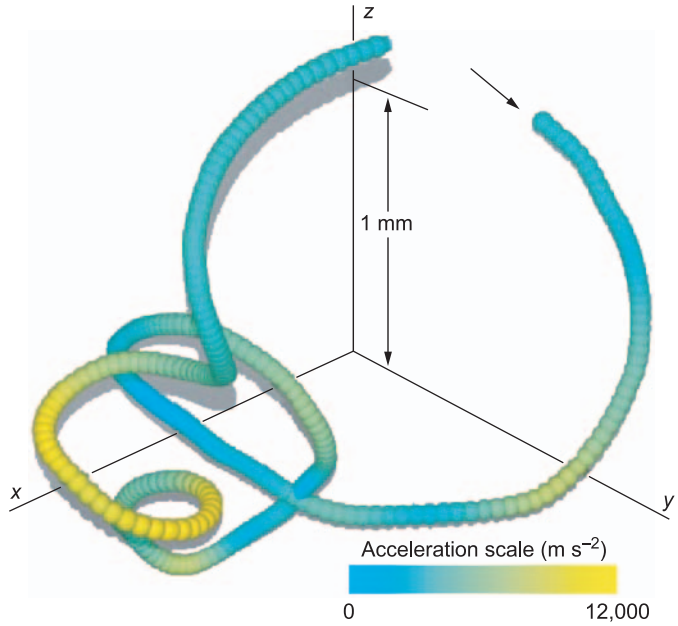


Figure 8. Three-Dimensional Particle Trajectory in a Turbulent Fluid
A high-speed silicon-strip detector was used to record this trajectory of a particle in a turbulent fluid with *Re* = 63,000 (La Porta 2001). The magnitude of the instantaneous acceleration is color-coded. Averaging over many such trajectories allows comparison with the theory of Lagrangian acceleration statistics.
(Modified with permission from *Nature*. This research was performed at Cornell University.)

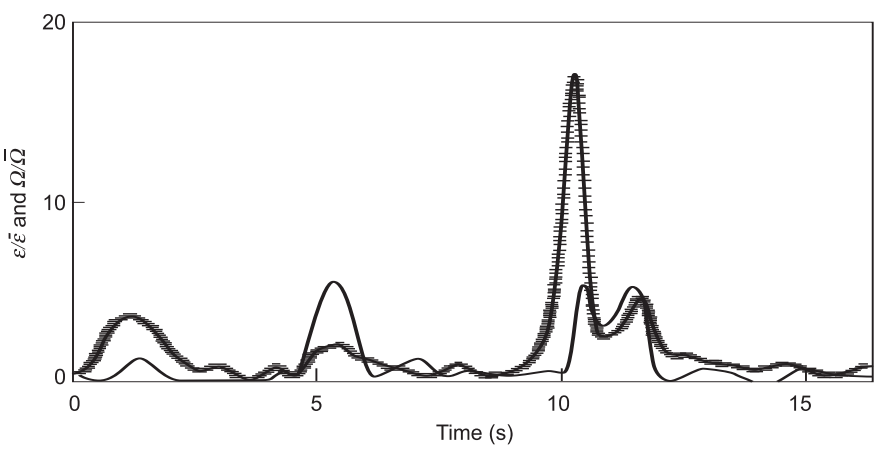


Figure 9. Intermittency of Energy Dissipation and Enstrophy at *Re* = 48,000
Time traces of the local energy dissipation ε (crosses) and enstrophy $\Omega = \langle \omega^2 \rangle/2$ (solid curve) illustrate the very intermittent behavior of these dissipation quantities for turbulent flow with *Re* = 48,000 (Zeff 2003). (Modified with permission from *Nature*. This research was performed at the University of Maryland.)

Finally, for physical realizations of 2-D flows, full-velocity fields can be measured with high resolution in both space and time (Rivera et al. 2003), and the 2-D velocity gradient tensor can be used to identify topological structures in the flow and correlate them with turbulent cascade mechanisms. The technique used to make these measurements and those represented in Figure 10 is particle-image velocimetry (PIV) or its improved version, particle-tracking velocimetry (PTV). Two digital images, taken

[10] The velocity gradient tensor is a 3×3 matrix consisting of the spatial derivatives of three components of velocity. For example, for the velocity component u_i, the derivatives are $\partial u_i/\partial x_1$, $\partial u_i/\partial x_2$, and $\partial u_i/\partial x_3$.

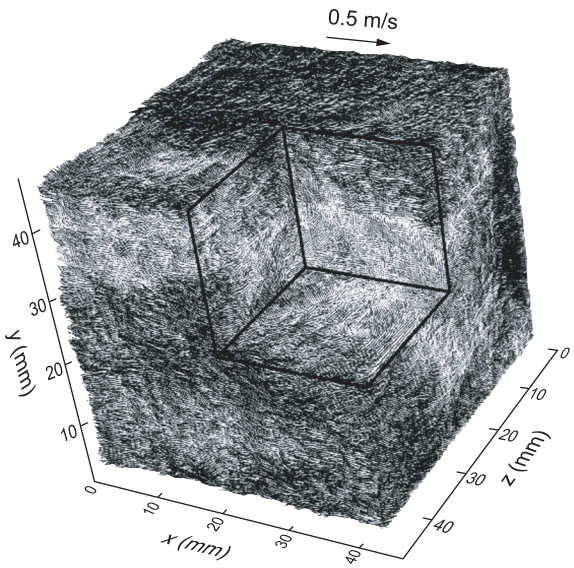


Figure 10. High-Resolution 3-D Turbulent Velocity Fields
These images were obtained using digital holographic particle-imaging velocimetry.
(Zhang et al. 1997. Modified with permission from *Experiments in Fluids*.)

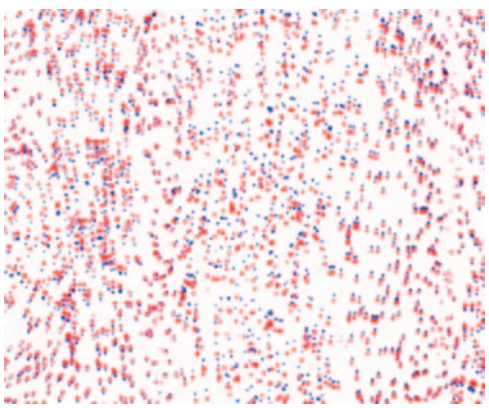


Figure 11. Superimposed Digital Images of a Particle Field
Two digital images of suspended particles taken 0.003 s apart are superimposed. The first exposure is in red; the second, in blue. In PIV, the pattern of particles over a small subregion is correlated between exposures, and an average velocity is computed by the mean displacement δx of the pattern. In PTV, each particle is matched between exposures; as a result, spatial resolution is higher, and there is no spatial averaging. These data allow one to infer the velocity field connecting the two images.

closely spaced in time, track the motion of small particles that seed the flow and move with the fluid. The basic notion is illustrated in Figure 11, where two superimposed digital images of particle fields, separated by $\Delta t = 0.03$ second, are shown. Within small subregions of the domain, patterns of red particles in image 1 can be matched with very similar patterns of blue particles in image 2 by maximizing the pattern correlation. An average velocity vector for the matching patterns is then calculated over the entire box from which a velocity field is obtained, as shown in Figure 12(a). Notice that there are some anomalous vectors caused by bad matching that need to be fixed by some interpolation scheme. PTV, on the other hand, uses a particle-matching algorithm to track individual particles between frames. The resulting vector field is shown in Figure 12(b). The PTV method has higher spatial resolution than PIV but also greater computational-processing demands and more stringent image-quality constraints. From the PTV velocity field, the full vorticity field $\boldsymbol{\omega}$ can be computed as shown in Figure 12(c).

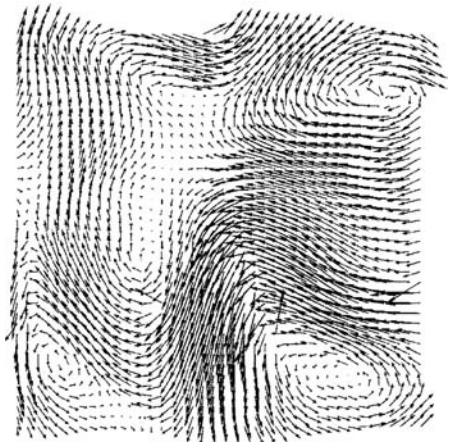
An additional advantage of the PTV approach is that, for high enough temporal resolution, individual particle tracks can be measured over many contiguous frames, and information about Lagrangian particle trajectories can be obtained. Some 2-D particle tracks are shown in Figure 13.

This capability can be combined with new analysis methods for turbulence to produce remarkable new visualization tools for turbulence. Figure 14 shows the full backward and forward time evolution of a marked region of fluid within an identified stable coherent structure (a vortex). These fully resolved measurements in two dimensions will help build intuition for the eventual development of similar capabilities in three dimensions. Further, the physical mechanisms of 2-D turbulence are fascinating in their own right and may be highly relevant to atmospheric or oceanic turbulence.

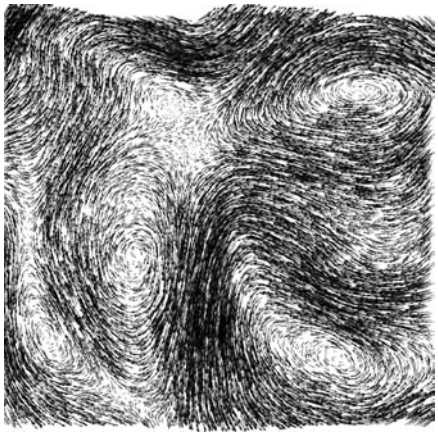
The Prospects for "Solving" Turbulence

Until recently, the study of turbulence has been hampered by limited experimental and numerical data on the one hand and the extremely intractable form of the Navier-Stokes equation on

(a)



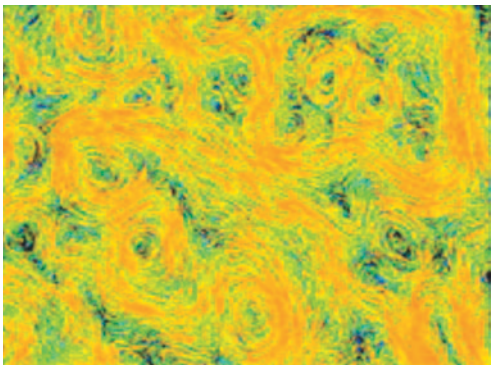
(b)



(c)

Figure 12. Two-Dimensional Vector Velocity and Vorticity Fields
The velocities in (a) and (b) were obtained from data similar to those in Figure 11 using PIV and PTV techniques, respectively. The vorticity field in (c) is calculated from the velocity field in (b).

(a)



(b)

Figure 13. Individual Lagrangian Particle Tracks for Forced 2-D Turbulence
(a) Approximately 10^4 particles are tracked for short periods.
(b) Several individual trajectories are shown for several injection-scale turnover times.

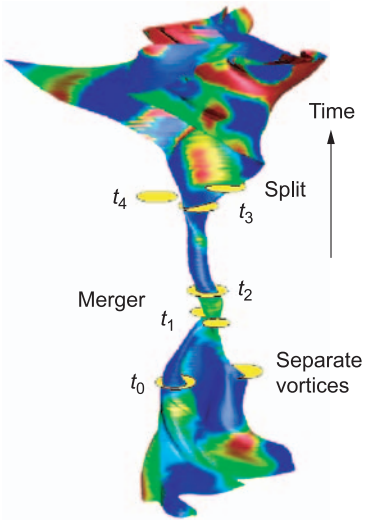


Figure 14. Time Evolution of a Compact Distribution of 10^4 Points in a Coherent Vortex at Time t_2
Vortex merging and splitting happen at times t_1 and t_3, respectively. The color-coding of the surface represents the spatially local energy flux to larger (red) and smaller (blue) spatial scales.
(M. K. Rivera, W. B. Daniel, and R. E. Ecke, to be published in Gallery of Fluid Images, Chaos, 2004.)

the other. Today, the advent of large-scale scientific computation, combined with new capabilities in data acquisition and analysis, enables us to simulate and measure whole velocity fields at high spatial and temporal resolutions. Those data promise to revolutionize the study of fluid turbulence. Further, new emerging ideas in statistical hydrodynamics derived from field theory methods and concepts are providing new theoretical insights into the structure of turbulence (Falkovitch et al. 2001). We will soon have many of the necessary tools to attack the turbulence problem with some hope of solving it from the physics perspective if not with the mathematical rigor or the extremely precise prediction of properties obtained in, say, quantum electrodynamics. Let me explain then what I mean by that solution. In condensed matter physics, for example, the mystery of ordinary superconductivity was solved by the theory of Bardeen, Cooper, and Schrieffer (1957), which described how electron pairing mediated by phonons led to a Bose-Einstein condensation and gave rise to the superconducting state. Despite this solution, there has been no accurate calculation of a superconducting transition temperature for any superconducting material because of complications emerging from material properties. I think that there is hope for understanding the mechanisms of turbulent energy, vorticity, and mass transfer between scales and between points in space. This advance may turn out to be elegant enough and profound enough to be considered a solution to the mystery of turbulence. Nevertheless, because turbulence is probably a whole set of problems rather than a single one, many aspects of turbulence will likely require different approaches. It will certainly be interesting to see how our improved understanding of turbulence contributes to new predictability of one of the oldest and richest areas in physics. ■

Acknowledgments

Three important books on fluid turbulence, McComb (1990), Frisch (1995), and Lesieur (1997), published in the last 20 years have helped provide the basis for this review. I would like to thank Misha Chertkov, Greg Eyink, Susan Kurien, Beth Wingate, Darryl Holm, and Greg Swift for valuable suggestions. I am indebted to my scientific collaborators—Shiyi Chen, Brent Daniel, Greg Eyink, Michael Rivera, and Peter Vorobieff—for many exciting moments in learning about fluid turbulence, and I look forward to more in the future. I would also like to acknowledge the LDRD program at Los Alamos for its strong support of fundamental research in turbulence during the last 5 years

Further Reading

Bardeen, J., L. N. Cooper, and J. R. Schrieffer. 1957. Theory of Superconductivity. *Phys. Rev.* **108** (5): 1175.

Batchelor, G. K. 1952. The Effect of Homogeneous Turbulence on Material Lines and Surfaces. *Proc. R. Soc. London, Ser. A* **213**: 349.

———. 1953. *The Theory of Homogeneous Turbulence*. Cambridge, UK: Cambridge University Press.

———. 1959. Small-Scale Variation of Convected Quantities Like Temperature in Turbulent Fluid. Part 1. General Discussion and the Case of Small Conductivity. *J. Fluid Mech.* **5**: 113.

———. 1969. Computation of the Energy Spectrum in Homogeneous Two-Dimensional Turbulence. *Phys. Fluids* **12**, Supplement II: 233.

Besnard, D., F. H. Harlow, N. L. Johnson, R. Rauenzahn, and J. Wolfe. 1987. Instabilities and Turbulence. *Los Alamos Science* **15**: 145.

Champagne, F. H. 1978. The Fine-Scale Structure of the Turbulent Velocity Field. *J. Fluid Mech.* **86**: 67.

Falkovitch, G., K. Gawedzki, and M. Vergassola. 2001. Particles and Fields in Fluid Turbulence. *Rev. Mod. Phys.* **73**: 913.

Frisch, U. 1995. *Turbulence*: *The Legacy of A. N. Kolmogorov*. Cambridge, UK: Cambridge University Press.

Heisenberg, W. 1948. Zur Statistischen Theorie der Turbulenz. *Z. Phys.* **124**: 628.

Kaneda, Y., T. Ishihara, M. Yokokawa, K. Itakura, and A. Uno. 2003. Energy Dissipation Rate and Energy Spectrum in High Resolution Direct Numerical Simulations of Turbulence in a Periodic Box. *Phys. Fluids* **15**: L21.

Kolmogorov, A. N. 1941a. The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Number. *Dok. Akad. Nauk. SSSR* **30**: 9

———. 1941b. Decay of Isotropic Turbulence in an Incompressible Viscous Fluid. *Dok. Akad. Nauk. SSSR* **31**: 538.

———. 1941c. Energy Dissipation in Locally Isotropic Turbulence. *Dok. Akad. Nauk. SSSR* **32**: 19.

Kraichnan, R. H. 1959. The Structure of Isotropic Turbulence at Very High Reynolds Numbers. *J. Fluid Mech.* **5**: 497.

———. 1965. Lagrangian-History Closure Approximation for Turbulence. *Phys. Fluids* **8**: 575.

———. 1967. Inertial Ranges in Two-Dimensional Turbulence. *Phys. Fluids* ***10***: 1417.

———. 1968. Small-Scale Structure of a Scalar Field Convected by Turbulence. *Phys. Fluids* **11**: 945.

———. 1974. On Kolmogorov's Inertial-Range Theories. *J. Fluid Mech.* **62**: 305.

Lamb, C. 1932. *Address to the British Association for the Advancement of Science. Quoted in Computational Fluid Mechanics and Heat Transfer* by J. C. Tannehill, D. A. Anderson, and R. H. Pletcher. 1984. New York: Hemisphere Publishers.

La Porta, A., G. A. Voth, A. M. Crawford, J. Alexander, and E. Bodenschatz. 2001. Fluid Particle Accelerations in Fully Developed Turbulence. *Nature* **409**: 1017.

Lesieur, M. 1997. *Turbulence in Fluids*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

McComb, W. D. 1990. *The Physics of Fluid Turbulence*. Oxford, UK: Oxford University Press.

Meneveau, C., and J. Katz. 2000. Scale-Invariance and Turbulence Models for Large-Eddy Simulation. *Annu. Rev. Fluid Mech.* **32**: 1.

Obukhov, A. M. 1941. Spectral Energy Distribution in a Turbulent Flow. *Dok. Akad. Nauk. SSSR* **32**: 22.

Onsager, L. 1945. The Distribution of Energy in Turbulence. *Phys. Rev.* **68**: 286.

———. 1949. Statistical Hydrodynamics. *Nuovo Cimento* **6**: 279.

Orszag, S. A., and G. S. Patterson. 1972. In *Statistical Models and Turbulence*, *Lecture Notes in Physics*, Vol. **12**, p. 127. Edited by M. Rosenblatt and C. M. Van Atta. Berlin: Springer-Verlag.

Reynolds, O. 1895. On the Dynamical Theory of Incompressible Viscous Fluids and the Determination of the Criterion. *Philos. Trans. R. Soc. London*, *Ser. A* **186**: 123.

Richter, J. P. 1970. Plate 20 and Note 389. In *The Notebooks of Leonardo Da Vinci*. New York: Dover Publications.

Richardson, L. F. 1922. *Weather Prediction by Numerical Process*. London: Cambridge University Press.

———. 1926. Atmospheric Diffusion Shown on a Distance-Neighbour Graph. *Proc. R. Soc. London*, *Ser. A* **110**: 709.

Rivera, M. K., W. B. Daniel, S. Y. Chen, and R. E. Ecke. 2003. Energy and Enstrophy Transfer in Decaying Two-Dimensional Turbulence. *Phys. Rev. Lett.* **90**: 104502.

She, Z. S., S. Y. Chen, G. Doolen, R. H. Kraichnan, and S. A. Orszag. 1993. Reynolds-Number Dependence of Isotropic Navier-Stokes Turbulence. *Phys. Rev. Lett.* **70**: 3251.

Taylor, G. I. 1935. Statistical Theory of Turbulence. *Proc. R. Soc. London*, *Ser. A* **151**: 421.

———. 1938. The Spectrum of Turbulence. *Proc. R. Soc. London*, *Ser. A* **164**: 476.

Van Dyke, M. 1982. *An Album of Fluid Motion*. Stanford, CA: Parabolic Press.

von Kármán, T., and L. Howarth. 1938. On the Statistical Theory of Isotropic Turbulence. *Proc. R. Soc. London*, *Ser. A* **164**: 192.

von Neumann, J. 1963. Recent Theories of Turbulence. In *Collected Works* (1949–1963), Vol. 6, p. 437. Edited by A. H. Taub. Oxford, UK: Pergamon Press.

Zeff, B. W., D. D. Lanterman, R. McAllister, R. Roy, E. J. Kostelich, and D. P. Lathrop. 2003. Measuring Intense Rotation and Dissipation in Turbulent Flows. *Nature* **421**: 146.

Zhang, J., B. Tao, and J. Katz. 1997. Turbulent Flow Measurement in a Square Duct with Hybrid Holographic PIV. *Exp. Fluids* **23**: 373.

For further information, contact Robert Ecke (505) 667-6733 (ecke@lanl.gov).

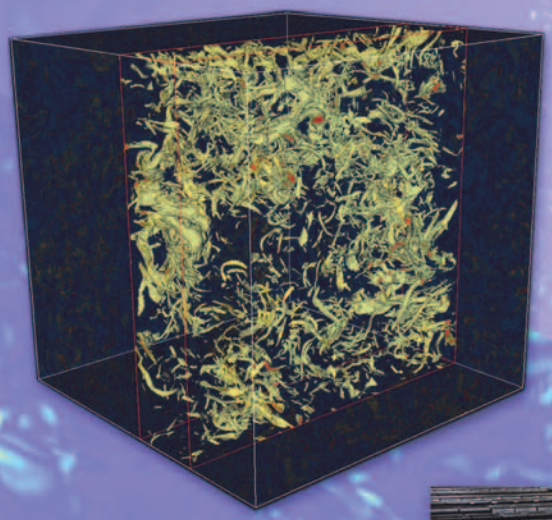
Direct Numerical Simulations of Turbulence

Data Generation and Statistical Analysis

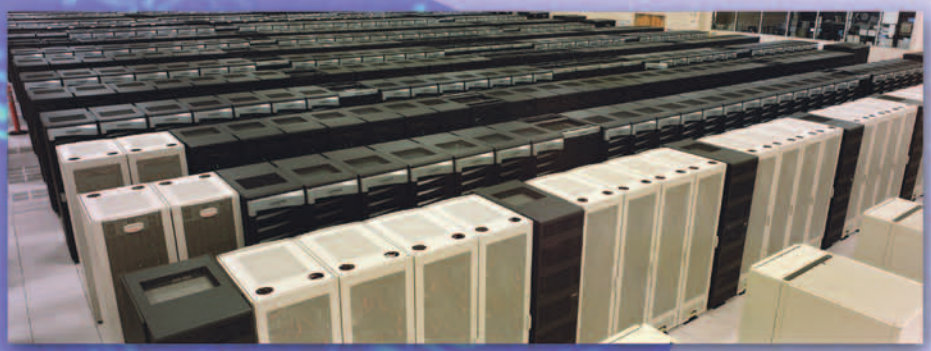
Susan Kurien and Mark A. Taylor

In 1941, Andrei N. Kolmogorov predicted that, within all highly turbulent flows, there is a universal energy-conserving cascade whereby the energy of the large-scale eddies is transferred to finer and finer scales, down to the scales at which the energy is finally dissipated to heat. It is difficult to measure such a cascade directly, but related benchmark predictions for the statistical behavior of turbulent flows can now be calculated and examined using advanced simulation and flow visualization tools. Los Alamos scientists have been able to simulate flows of Reynolds numbers up to 10^5, the largest of which needed of the order of terabytes of data storage and used the full power of the Advanced Simulation and Computing (ASC) Q machine for several weeks of computer time. Through clever analysis of single frames of the simulations, a great deal of information can be extracted to show that the original constraints for the Kolmogorov theory can be relaxed so that, in fact, his statistical predictions hold locally in time. Furthermore, scientists are able to measure new statistical quantities that demonstrate the conditions under which departures from Kolmogorov theory begin to occur. This type of statistical analysis of numerical data is setting the agenda for future research.

Visualization of vorticity in a portion of a 256^3 subdomain of the 2048^3 turbulence simulation performed on the ASC Q machine at Los Alamos.



The ASC Q machine.

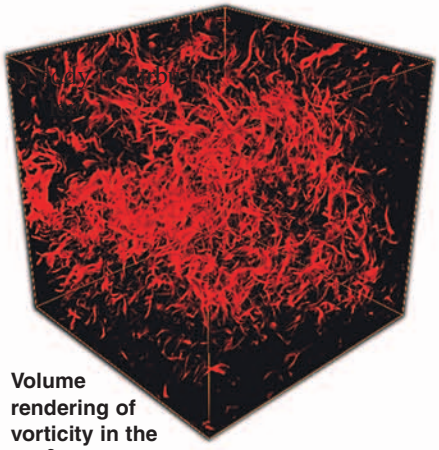


The problem of fluid turbulence has benefited from concerted efforts in theoretical, experimental, and most recently, computational research. However, while theoretical and experimental efforts have cooperated for some time to advance the field, computational science is a relatively recent entry and provides new data and problems that have not been accessible by more established techniques. For some problems, the entire turbulent flow field can now be calculated to high precision with suitable numerical methods. Flow visualization and extensive three-dimensional (3-D) statistical analysis, for example, are techniques that can be used profitably. Computational capabilities and expertise at Los Alamos National Laboratory have resulted in calculations that reveal new universal properties of turbulence and new directions in which to expand research efforts, as we describe below.

Solving the Navier-Stokes equations, which provide the best-known mathematical description of turbulent flow, remains an immensely challenging problem. However, turbulence research is driven by a practical need for real-world engineering applications and by the need to understand and predict the universal fundamental features, if any, in all turbulent phenomena. Therefore, approaches to studying turbulence other than computational ones have evolved over several decades and have produced a deep understanding of the subject on a fundamental as well as a phenomenological level. One such approach was initiated in the late 19th century by Osborne Reynolds, who proposed to ignore the details of the turbulent flow at each instant and, instead, to regard the flow as a superposition of mean and fluctuating parts. What naturally followed this shift in approach was the addition of statistics and probability theory to the arsenal of tools used to understand turbulence. The turbulence field is considered to be a random field in the probabilistic sense. The idea is to study the statistical moments of turbulent fields such as the multipoint correlation functions of velocity, pressure, and so on with the aim to recover the full probability-distribution function of the field and its evolution given a set of initial (boundary) conditions. Alternatively, there are attempts to obtain the probability distribution functions first and derive from them the statistics of the turbulent system. In a broad sense, deriving these functions is the goal of statistical hydrodynamics research (refer to the article " Field Theory and Statistical Hydrodynamics" on page 181 of this volume). This article will examine some of the questions that statistical analysis of turbulence data can address using several data sets generated by solving the Navier-Stokes equations on grids with different spatial resolutions.

Universal Properties of Turbulence

First, we briefly address the problem of universality of statistical properties. We would like to know whether turbulence exists independently of the type of flow (water flowing in a pipe or in a river, wind flow, and others), the fluid that is flowing (air or water), the boundary conditions (smooth, rough, artificial, or periodic), or the energy-input mechanisms (stirring, shaking, or shearing). Is there a regime of length scales that has quantifiable properties common to all turbulent flows? Two phenomenological ideas have been useful in addressing this question. The first was proposed by Lewis F. Richardson in the late 19th century and is consistent with our intuition from observing turbulence—the energy input at large scales is transferred into successively smaller eddies of the turbulent flow in a so-called cascade process. The notion of



Volume rendering of vorticity in the 256^3 subdomain shown on the opposite page.

an "eddy" in turbulent flow is somewhat nebulous, but for current purposes, it should be thought of as a coherent turbulence structure with an associated length scale, location, and lifetime. The second idea is a hypothesis advanced by Andrei N. Kolmogorov (1941): For highly turbulent flows in which the Richardson cascade has created many generations of eddies, the turbulent length scales of size r that are much smaller than the typical large scale L of the flow and much larger than the viscous dissipative scale η must have universal statistical properties. Kolmogorov conjectured that, in this regime of intermediate scales, the dynamics is minimally affected by forcing, boundaries, and large-scale anisotropies, which are generally flow-dependent, and unaffected by the viscous dissipative effects that occur at the very small scales. The dynamics in this so-called inertial range are dominated by the nonlinear term of the Navier-Stokes equations, and it seems reasonable that inertial-range dynamics should display universal behavior statistically. In our discussion of new statistical-analysis and diagnostic techniques, we will be concerned primarily with the statistics of this universal inertial range of scales in high-Reynolds-number turbulence (see the article " The Turbulence Problem" on page 124

for definitions of these terms).

The typical statistical scale-dependent quantities investigated are known as structure functions, one type of which is

$$S_n(r) = \left\langle \left[u_L(\mathbf{x}+\mathbf{r}) - u_L(\mathbf{x}) \right]^n \right\rangle , \quad (1)$$

where $u_L(x) = \mathbf{u}(x)\cdot\hat{\mathbf{r}}$ is the component of the velocity along $\mathbf{r}$ (the subscript L denotes longitudinal velocity) and $\langle ... \rangle$ denotes ensemble and domain averaging over all $\mathbf{x}$. This structure function is thus the nth-order moment of the velocity difference across scales of size r and is a measure, order by order, of the statistical properties of eddies of size r. Kolmogorov derived a fundamental physical law for the inertial range of scales r for high Reynolds number, slowly decaying (essentially steady-state) turbulence under the assumption of isotropy and homogeneity of the small scales:

$$S_3(r) = \left\langle \left[u_L(\mathbf{x}+\mathbf{r}) - u_L(\mathbf{x}) \right]^3 \right\rangle = -\frac{4}{5}\varepsilon r , \quad (2)$$

where ε is the mean rate of energy flux balancing the mean rate of energy dissipation in statistically steady turbulence in the limit of zero viscosity. This so-called "four-fifths law" (Kolmogorov 1941) is a statement of energy conservation in the inertial range; that is, the energy flux through scales of size r_1 equals the energy flux through scales of size r_2 if both r_1 and r_2 are in the inertial range. The four-fifths law is now used as a nominal measure of the regime of inertial-range scaling in experimental and numerical data; that is, the range of scales over which the four-fifths law is close to being satisfied is taken to be the statistically "universal" scaling regime.

Kolmogorov also assumed that the cascade of energy occurs in a space-filling, self-similar way. Formally, there exists a unique scaling exponent h such that

$$S_n(\lambda r) = \lambda^h S_n(r) \quad . \quad (3)$$

To be consistent with the four-fifths law, the assumption of self-similarity implies that $h = 1/3$ and that, in general, if structure functions of arbitrary order are to scale with r, then

$$S_n(r) \approx r^{\zeta_n}, \text{ where } \zeta_n = \frac{n}{3} \quad . \quad (4)$$

Most of the known empirical departures from the Kolmogorov scaling prediction can be traced to three causes: The Reynolds number is not large enough, the scaling is contaminated by the anisotropies inevitable in most flows, and the self-similarity assumption is not valid. The effects relating to small Reynolds numbers are something we have to live with, in a sense, because of the limitations of technology and computational power, but cumulative data analysis of experiments and simulations performed over several decades strongly suggest that the scaling exponents do not differ much for a Taylor microscale Reynolds number[1] R_λ ranging from approximately 100 to approximately 10,000.

It therefore seems that, at a minimum, we observe a convergence of the exponents over a wide range of high Reynolds numbers. The assumption of statistical isotropy, that is, invariance under arbitrary rigid rotations, is key to the scaling-law predictions, but isotropy is a rather strong restriction to make when most turbulent flows are apparently highly anisotropic. There are two ways to remove the inevitable effects of anisotropy in order to test the fundamental assumption of self-similarity. The first is to measure flows with extremely high Reynolds numbers, such as wind flow over the ground, that yield wide separation of scales and resort to the Kolmogorov assumption that, for sufficiently small scales, the statistics will be locally isotropic. The second is to explicitly extract the isotropic component of the statistics, for example, by systematically averaging out the anisotropic contributions, as we discuss in detail below. Recently, the effect of anisotropy on scaling exponents has been studied extensively, and there are now ways to quantify anisotropic effects (Kurien and Sreenivasan 2001), as well as to extract purely isotropic contributions (Taylor et al. 2003), which might then be more sensibly compared with theoretical predictions. We will discuss a new method to implement the latter procedure that has proved to be very useful in analyzing arbitrarily anisotropic flows. The final known reason for departure from the Kolmogorov scaling prediction is that the turbulent cascade is not self-similar. That is, instead of each generation of eddies being produced in a space-filling, self-similar way, the cascade proceeds in an intermittent manner, in which some parts of the flow at a given instant are extremely active while others are relatively quiescent. This is the now well-known intermittency feature of turbulence, and it results in what is known as "anomalous" scaling—that is, there is no unique scaling exponent h from which all scaling exponents can be simply derived.

In the remainder of this article, we describe our studies of the universal statistical features of turbulence using quantities such as the structure functions measured from simulations

[1]The Taylor microscale Reynolds number is $R_\lambda = u'\lambda/\nu$, where u' is the velocity fluctuation and ν is the viscosity. Initially, G. I. Taylor thought that the scale λ—the radius of curvature at the origin of the autocorrelation of the fluctuating velocity—was the viscous dissipation scale of turbulence. In fact, its magnitude is intermediate between the large scale L and the true (Kolmogorov) dissipation scale η. R_λ is often used instead of the large-scale Reynolds number, Re, to characterize flows that have widely varying large-scale properties and, hence, widely varying Reynolds numbers but whose small-scale fluctuations might be comparable. At high Reynolds numbers, $R_\lambda \propto Re^{1/2}$.

(resolved down to the dissipation scale) of the fundamental equations of motion, the Navier-Stokes equations. First, we discuss the simulations themselves and then demonstrate the use of diagnostics to extract statistically isotropic features of the flow. Our results suggest a refinement of the Kolmogorov picture of isotropic turbulence.

Direct Numerical Simulations

Direct numerical simulation (DNS) refers to solving the Navier-Stokes equations numerically by resolving all scales down to the scale of viscous dissipation. DNS represents a brute-force approach to modeling turbulence: No modeling is required beyond the Navier-Stokes equations, simple well-understood numerical methods are used, but massive computing resources are needed. When carefully produced, DNS data is an excellent substitute for exact, analytic solutions of the Navier-Stokes equations. The only drawback is that to obtain solutions for moderately high Reynolds numbers requires weeks of computing time on today's largest supercomputers. To achieve the Reynolds numbers of a typical atmospheric boundary layer flow, $R_\lambda = 10{,}000$, will require a 10^8-fold increase in computing power over today's largest computers. Fortunately, large-scale features such as the mean flow and other statistical properties of turbulence depend only weakly on the Reynolds number. Thus, DNS of flows with more moderate Reynolds numbers has been valuable for studying many aspects of turbulence, including universal statistical features. For additional information, see, for example, the review by Moin and Mahesh (1998).

To obtain as high a Reynolds number as possible, DNS calculations are usually performed on the simplest flows: the incompressible Navier-Stokes equations, without multiple materials or other physics that must be modeled. The calculations are further limited to simple domains and equally spaced grids, which allow for very efficient numerical algorithms. The highest possible Reynolds numbers can be achieved for the classic problem of homogeneous turbulence in a square box with periodic boundary conditions, the problem we have focused on.

For fully resolved calculations, spectral methods are preferred for their high accuracy. Although high-order finite-difference codes can yield similar accuracy, spectral methods still have an advantage because they permit fast, direct solution of Poisson's equation. Solving Poisson's equation is required to determine the pressure gradient that appears in the Navier-Stokes equations. Spectral methods became practical for computational fluid dynamics after the development of the spectral-transform method (Eliasen et al. 1970, Orszag 1970). Additional issues important for the Navier-Stokes equations, such as time-stepping schemes and control of aliasing errors, were effectively treated in Rogallo (1981). The methods used today are quite similar to those used in that work.

The spectral part of a DNS code refers to the method used for the spatial discretization of the equations. In particular, to compute a spatial derivative of a term in the equations, one first expands that term in a truncated Fourier expansion using the fast Fourier transform (FFT) and then computes the derivatives exactly from the Fourier expansion. After the equations are discretized in space, we are left with a system of ordinary differential equations, which are integrated in time with a third- or fourth-order Runge-Kutta or similar scheme. This procedure has one complication arising from the nonlinear advection term. The nonlinearity can transfer energy into frequencies higher than can be resolved by the numerical grid. The energy in these unresolved frequencies will then artificially contaminate the energy and phases of the resolved frequencies in a procedure known as aliasing. This aliasing error is typically controlled by properly designed spectral filters.

The computational expense of DNS comes from the strict restrictions on the grid spacing, Δx, and the time step, Δt, that are required to fully resolve all scales in the Navier-Stokes equations. If one is primarily interested in the statistical properties of the inertial range, it is sufficient to run the numerical simulation with $\Delta x \leq 3\eta$, where η is the Kolmogorov length scale.[2]

Since $\eta \sim Re^{-3/4}$ (or $\eta \sim R_\lambda^{-3/2}$), this grid-spacing restriction also determines the highest-Reynolds-number flow that can be accurately computed for a given Δx. The restrictions on Δt can be estimated from the consideration of physical time scales in the problem, but in practice, a more restrictive constraint comes from the CFL (Courant-Friedrichs-Levy) condition, which shows that, for the time-stepping schemes used, the time step must be kept proportional to the grid spacing. Combined, these considerations show that the computational cost of DNS is proportional to R_λ^6 (Pope 2000).

In DNS calculations, it is important to ensure that the energy dissipation is due entirely to the viscous terms in the Navier-Stokes equations, rather

[2] The Kolmogorov length scale η depends only on the rate of energy flux ε and the (chosen) fluid viscosity ν. In the forced simulations, η is determined entirely by the forcing (rate of input of energy), which balances the flux rate in the statistical steady state and the chosen viscosity coefficient. In the decaying simulation, η is fully determined at initial time by the initial condition but thereafter evolves with the dynamics, thus resulting in increasing resolution as the flow decays.

than to the numerical method used. Often, numerical methods are designed to introduce various types of artificial dissipation, which can have beneficial properties but are not appropriate for DNS. For the spectral method outlined here, we estimate the numerical viscosity by computing the kinetic energy E at every time step and comparing the numerical evolution of E,

$$\frac{dE}{dt} = \frac{E(t + \Delta t) - E(t)}{\Delta t} \quad ,$$

where $E = 0.5 \langle \mathbf{u} \cdot \mathbf{u} \rangle$, with the evolution given by the Navier-Stokes equations. In the unforced case, the latter term is

$$\frac{dE}{dt} = -\nu \langle \nabla \mathbf{u} \cdot \nabla \mathbf{u} \rangle \quad ,$$

where $\mathbf{u}$ is the flow field. In our largest simulation, the two quantities agree to more than four digits, demonstrating that over 99.99 percent of the dissipation is due to the Navier-Stokes viscosity.

Finally, if DNS in a periodic box is used to study universal features of turbulence, the largest scales are strongly influenced by the square computational domain. For example, consider a field with all its energy in spherical wave numbers of at most 2. There are only a handful of such Fourier modes, and they are strongly aligned with the coordinate directions of the box. Any such field could not be isotropic. Many of the directional moments of the field would greatly differ between coordinate and noncoordinate directions. There are several ways to avoid this effect in order to obtain more isotropic simulations. The most direct method is to simply keep energy out of the large scales. This is the approach usually taken with decaying turbulence simulations. For forced simulations, it is possible to achieve flows with much higher Reynolds numbers by injecting energy into only the low wave numbers, but to obtain isotropic solutions requires careful attention. One approach is to use stochastic forcing designed so that the flow will be isotropic for a long enough time average, even though the field at any given time will have large anisotropies at the large scales. This approach introduces a lot of fluctuations in the solutions, so long time averages must be taken to obtain converged statistics. The most efficient approach is to use smooth, deterministic low-wave-number forcing. Converged statistics can then be obtained with shorter time averages, but some anisotropy will persist throughout the flow. For many quantities of interest, however, this anisotropy can be removed with the angle-averaging techniques described below.

Table I. Parameters for Five DNS Datasets

Data Set	Grid Points (N)[1]	Forced Flow	R_λ	Duration[2]	Storage per Frame (GB)
1	512	Yes (deterministic)	250	7	3
2	512	Yes (stochastic)	250	12	3
3	512	Yes (deterministic with helicity input)	250	10	3
4	1024	Yes (deterministic)	460	3	24
5	2048	No (initial condition of Johns Hopkins experiment)	170	3	192

[1] N is the number of points on each side of the cubic computational grid.

[2] The duration of a run is measured in units of the large eddy turnover time.

In our work, we have examined DNS simulations for decaying turbulence, stochastically forced turbulence, and deterministically forced turbulence (refer to Table I). For the decaying turbulence simulations, a properly chosen initial condition is allowed to decay through the effects of viscosity. For the forced problems, the simulations are run until the forcing and dissipation reach statistical equilibrium, and then they are run for several additional eddy turnover times to collect data from the equilibrium regime.

The decaying problem has the advantage that more realistic flows can be simulated, and it is possible, in principle, to compare the simulation results with those from experiments, such as those carried out at the recently upgraded Corrsin Wind Tunnel (Kang et al. 2003). But the decaying problem has the drawback that the results strongly depend on the initial condition, and one is faced with the challenge of generating a realistic turbulence state to use for the initial condition. To address this problem, in data set 5, we have followed the procedure described by Kang et al. (2003). We generate an initial flow field with random, uncorrelated phases but a prescribed energy spectrum. The flow is then run for a short time, until the phases become correlated enough to give a reasonable mean-derivative skewness. The energy spectrum is then reinitialized back to the original spectrum while retaining the correlated phases. Our low-wave-number forcing schemes are described in detail in Taylor et al. (2003). The deterministic forcing is based on the work by Sullivan et al. (1994), Sreenivasan et al. (1996), and Overholt and Pope (1998). Data sets 1 and 4 were obtained with this forcing. Data set 3 was obtained with a similar

scheme, but modified to inject helicity into the flow. The stochastic forcing used for data set 2 was based on the forcing given in Gotoh et al. (2002). We used both types of forcing to demonstrate the equivalence of the results when angle averaging is applied to the data.

Parallel Computing Issues

DNS calculations at resolutions of up to 512^3 can now be obtained on moderately large clusters. But the larger DNS calculations currently require Advanced Simulation and Computing (ASC)-class supercomputers. Our largest simulation, with a resolution of 2048^3, requires a 256-fold increase in computing power over that required for a resolution of 512^3. With 8 billion grid points, our 2048^3 simulation is one of the largest ever completed. It required several weeks using 2048 processors of ASC-Q and was made as part of the Laboratory's Science Runs to showcase ASC-Q's performance.

To implement FFT-based DNS codes on distributed memory parallel computers, the community relies almost exclusively on the data-transpose method. Each processor must perform thousands of FFTs per time step, but the data required for those FFTs will be distributed among many other processors. It is quite difficult to write an efficient, distributed-data FFT, and thus the data-transpose method continuously adjusts the distribution of data among the processors so that each processor can use a conventional serial FFT. The name "transpose" comes from the fact that if the data distribution is represented on a 3-D mesh of processors, the operations required by the data-transpose algorithm look like matrix transposes. For a resolution of 2048^3, over a terabyte of data must be moved through the network for each time step, and thus the method relies on a tightly coupled parallel computer with very high bandwidth. On ASC-Q, for problems of size N^3, we obtain excellent scaling for up to $N/2$ processors. Using N processors still represents a significant speedup, but the scalability starts to decrease, so there is little benefit to using more than N processors.

Another important problem concerns data input/output (I/O). For a resolution of 2048^3, each flow snapshot (which can also be used as a restart file) is 192 gigabytes. Serial I/O (having a single processor collect the data from all other processors and write it into a single file) can obtain data rates only in the tens of megabytes per second and thus requires hours to write a single snapshot or read in a snapshot when restarting. To avoid this unacceptable bottleneck, we utilized the Unified Data Model (UDM) I/O library of the High-Performance Computing Environment Group at Los Alamos. UDM, in conjunction with ASC-Q's parallel-file system, allows all processors to participate in the I/O for a single file. With UDM, we were able to obtain data-transfer rates of over 500 megabytes per second, which means snapshots can be written or read in under 7 minutes.

The Angle-Averaging Technique

In general, the two-point structure function $S(r)$ defined in Equation (1) is a function of the vector $\mathbf{r}$, that is, a function of the size of the separation scale $r = |\mathbf{r}|$, as well as of the orientation of $\mathbf{r}$. The Kolmogorov 1941 theory, however, assumes that, for sufficiently small scales, the flow depends only on the magnitude of $\mathbf{r}$ and is independent of the orientation of $\mathbf{r}$. Most reasonably controlled flow experiments (for example, those occurring in wind tunnels or pipes), as well as uncontrolled experiments (for example, those involving measurements of velocity in the atmospheric boundary layer), inevitably have some degree of anisotropy either from boundary configurations or from forcing mechanisms. Therefore, reasonable comparisons with theoretical predictions require understanding the degree of contamination caused by arbitrary anisotropy as well as formulating methods to eliminate these effects from the data. From experiments at very high Reynolds numbers (Taylor Reynolds number of ~10,000 or higher), in which there is wide separation between the large scales and the dissipative scales, we know that, for two-point statistics of the structure functions given in Equation (1), the contamination due to anisotropy decays rapidly with scale size and that local isotropy is recovered in the leading order. In numerical simulations, the Reynolds numbers, as well as the range of scales computed, are much smaller, and anisotropic effects typically do not have enough range of scales to decay sufficiently. As a result, they have a significant contribution in the inertial range. However, the availability of the full spatial and temporal information of the flow field offers other unique possibilities for investigating purely isotropic effects. One general procedure recently developed at Los Alamos is the angle averaging of the structure functions, which averages out the anisotropic contributions of an arbitrary (anisotropic) flow.

The primary motivation for our angle-averaging procedure is the recent derivation of a new version of the Kolmogorov four-fifths law (Duchon and Robert 2000, Eyink 2003). In this version, the four-fifths law states that for any domain B of size R in the limit that the viscosity $\nu \to 0$ (infinite or sufficiently high

Reynolds number), for scales of size $r << R$, and at any instant in time,

$$S_3(r) = \left\langle \left[u_L(\mathbf{x}+\mathbf{r}) - u_L(\mathbf{x}) \right]^3 \right\rangle$$
$$= \int \frac{d\Omega_r}{4\pi} \int d\mathbf{x} \left[u_L(\mathbf{x}+\mathbf{r}) - u_L(\mathbf{x}) \right]^3 \quad (5)$$
$$= -\frac{4}{5} \varepsilon_B r \ ,$$

where ε_B is the energy dissipation rate averaged over B. That is, the four-fifths law holds locally, instantaneously, and without any assumption of homogeneity or isotropy. The integration over the solid angle Ω, indicates averaging over all possible orientations of $\mathbf{r}$ for a given $|\mathbf{r}|$, which projects out the isotropic part of the correlation. The statement of energy conservation in the inertial range is now quite different—there is an underlying isotropic component common to all flows that formally obeys the same law that Kolmogorov derived using more restrictive assumptions.

To test this prediction with numerical simulations, we devised a way to take the solid-angle average of the data computed on a grid. The obvious, but computationally expensive and error-prone solution, would be to interpolate the velocity vector field over a sphere of desired radius r and integrate. Instead, we chose to first use the separation vectors allowed by the grid to compute structure functions for a fixed (θ,φ) as a function of r, as follows:

$$S_3(r,\theta,\phi) = \int d\mathbf{x} \left[u_L(\mathbf{x}+\mathbf{r}) - u_L(\mathbf{x}) \right]^3 \ .$$

Then, we computed a set of these structure functions for various (θ,φ) allowed by our grid so that we have a set of functions $S(r,\theta_1,\varphi_1)$, $S(r,\theta_2,\varphi_2),\ldots$ $S(r,\theta_n,\varphi_n)$ for pairs of angles (θ_i,φ_i) that span the full spherical solid angle rather uniformly. Each S is now a smooth function of r in a particular direction and can be interpolated to obtain $S(r)$ for any r. Then, to yield the angle-averaged value for a particular r, we compute

$$S_3(r) = \frac{1}{n} \sum_{i=1}^{n} w_i S_3(r,\theta_i,\phi_i) \ . \quad (6)$$

where the weight w_i is the solid angle subtended by the Voronoi cell containing the point $\hat{\mathbf{r}}$.

As $n \to \infty$, the average becomes arbitrarily close to the true spherical integral of Equation (5), and so the isotropic component of the statistics is recovered. The method is not specific to the four-fifths law and can in principle be used to examine the underlying isotropic component of any two-point correlation function, as we demonstrate below.

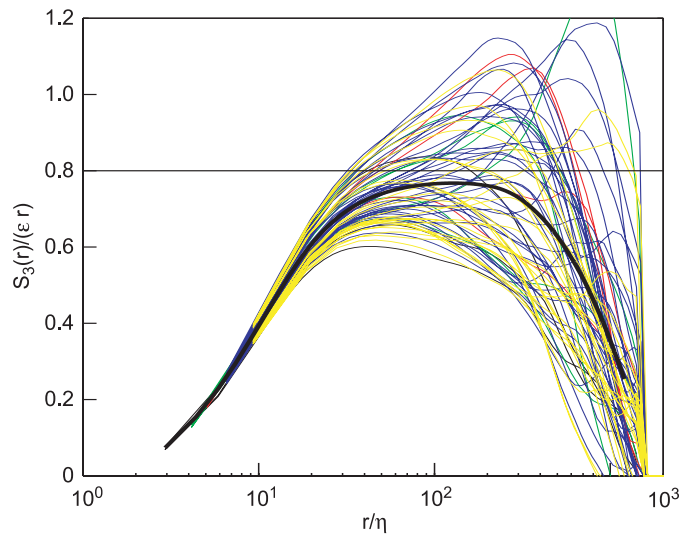


Figure 1. The Four-Fifths Law for a Single Frame of Forced Flow
The four-fifths law was computed for a single frame of data set 4 for deterministic forced flow, whose resolution is 1024^3. Each colored line is the compensated third-order structure function computed in one of 73 different directions of the flow. The black line is the angle-averaged function, which displays a range of scales between 30 and 200 that fall within 5% of the theoretical value of 0.8.

The Four-Fifths Law

Figure 1 shows such a calculation performed on a single frame of an anisotropically forced flow at a resolution of 1024 grid points to a side with periodic boundary conditions (data set 4), which was run long enough to achieve a statistically steady state. Each colored line is the compensated, domain-averaged, longitudinal third-order structure function, $S_3(r)/\varepsilon r$, computed in a particular direction in the periodic box for the increments r allowed by the grid in that direction. The length scale r has been nondimensionalized by the dissipation length scale η. The compensated statistics were computed in 73 different directions that were fairly evenly distributed over the sphere. As is clearly seen, the calculation in a given direction yields a smooth curve, which we interpolated using a cubic spline fit to obtain $S_3(r)/\varepsilon r$ for arbitrary length r in a given direction. The different directions also clearly display a large degree of variability with respect to each other, which appears to diminish as the scales get very small but is significant in a midrange of scales wherein the inertial range might be expected to lie. The thick black line is the average over all 73 directions of $S_3(r)/\varepsilon r$ as a function of r calcu-

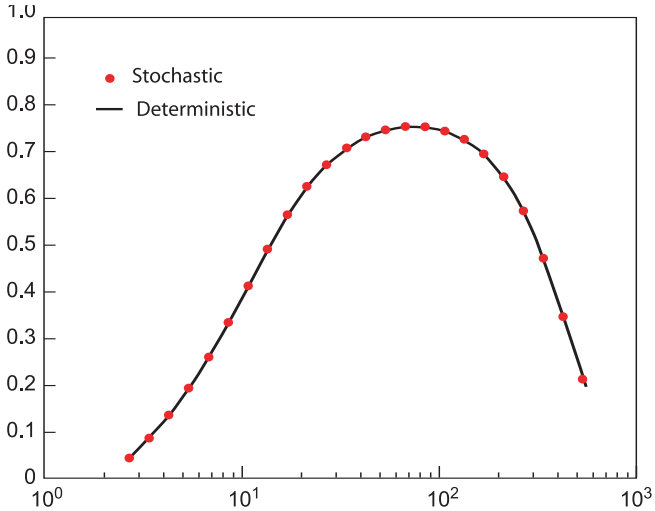


Figure 2. Angle- and Time-Averaged Compensated Third-Order Structure Function for Two Different Forced Flows

The angle- and time-averaged compensated third-order structure function was computed for data sets 1 (solid line) and 2 (dotted line), each of which has a resolution of 512^3. These two differently forced flows essentially coincide with each other in this statistical measure, thus supporting the notion of underlying universality of turbulent flows.

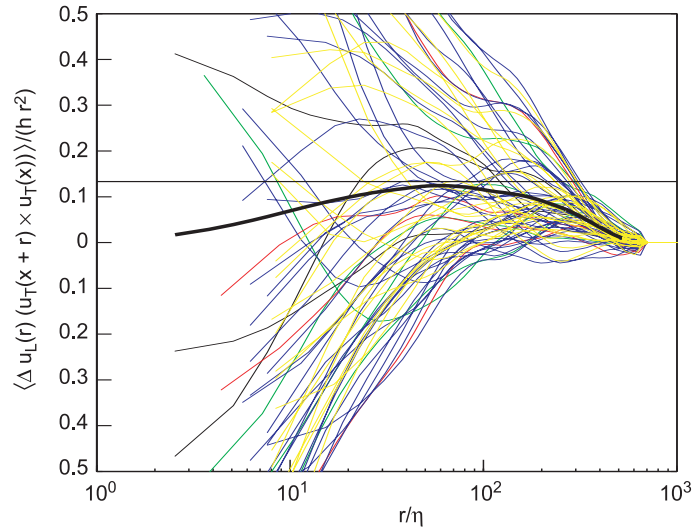


Figure 3. The Two-Fifteenths Law from a Single Frame of Data Set 3

The two-fifteenths law was computed for a single frame of data set 3, whose resolution is 512^3. Each colored line is the compensated third-order statistic in one of 73 different directions in the flow. The black curve is the angle-averaged function, which shows a range between 30 and 200 wherein its value is within 4% of the theoretically predicted value of 2/15.

lated according to Equation (6).

Remarkably, this angle-averaged function displays a reasonable range over which the curve is rather flat (indicating linear scaling in r) and is within 5 percent of 0.8, which is the theoretical predicted value. This result says that, at every instant in an anisotropic flow, there is an underlying isotropic component that can be projected out when an approximated spherical average is used and that, furthermore, obeys to a very good degree the fundamental universal four-fifths law for isotropic flow.

In Figure 2, we show the same calculation for data sets 1 and 2, which were calculated at lower Reynolds numbers but are forced in the low wave numbers as described above. The solid (black) and dotted (red) lines are the angle-averaged and then time-averaged compensated third-order structure functions for data sets 1 and 2, respectively. While the scaling range for this resolution is less than that in Figure 1, the noteworthy feature is that the curves are indistinguishable, which is a strong indication of universality because the underlying isotropic contributions of these two very different anisotropic flows are identical (Taylor et al. 2003).

The Two-Fifteenths Law

To demonstrate the distinction between the Kolmogorov local isotropy assumption and what we see in Figure 1, we discuss the measurement, using the same angle-averaging technique, of a very different statistical quantity that obeys the so-called two-fifteenths law:

$$\left\langle \left[u_L(\mathbf{x}-\mathbf{r}) - u_L(\mathbf{x})\right]\left[u_T(\mathbf{x}+\mathbf{r}) \times u_T(\mathbf{x})\right]\right\rangle = \frac{2}{15} h r^2 \quad , \qquad (7)$$

where h is the mean helicity dissipation rate and u_T denotes the component of $\mathbf{u}(\mathbf{r})$ transverse to $\mathbf{r}$. The quantity on the left side of this equation is a third-order statistic, as is $S_3(r)$ for the four-fifths law, but this new quantity probes the presence of a constant total helicity flux h in the inertial range (Kurien 2003). Like energy, helicity ($\mathbf{u} \cdot \nabla \times \mathbf{u}$) is conserved in turbulence, and our analysis has revealed that in the inertial range, helicity has other conserved properties in common with those of energy, such as constant flux.

Figure 3 shows this parity-breaking third-order statistic normalized by hr^2 in a forced flow in a periodic box of 512 grid points to a side with fixed sign of helicity input into the two lowest modes at each time step (data set 3). The picture in Figure 3 indicates that helicity flux (that is, the appropriate third-order correlation function) is highly anisotropic all the way into the small scales, as shown by the vast spread among the different directions. Nevertheless, there is still

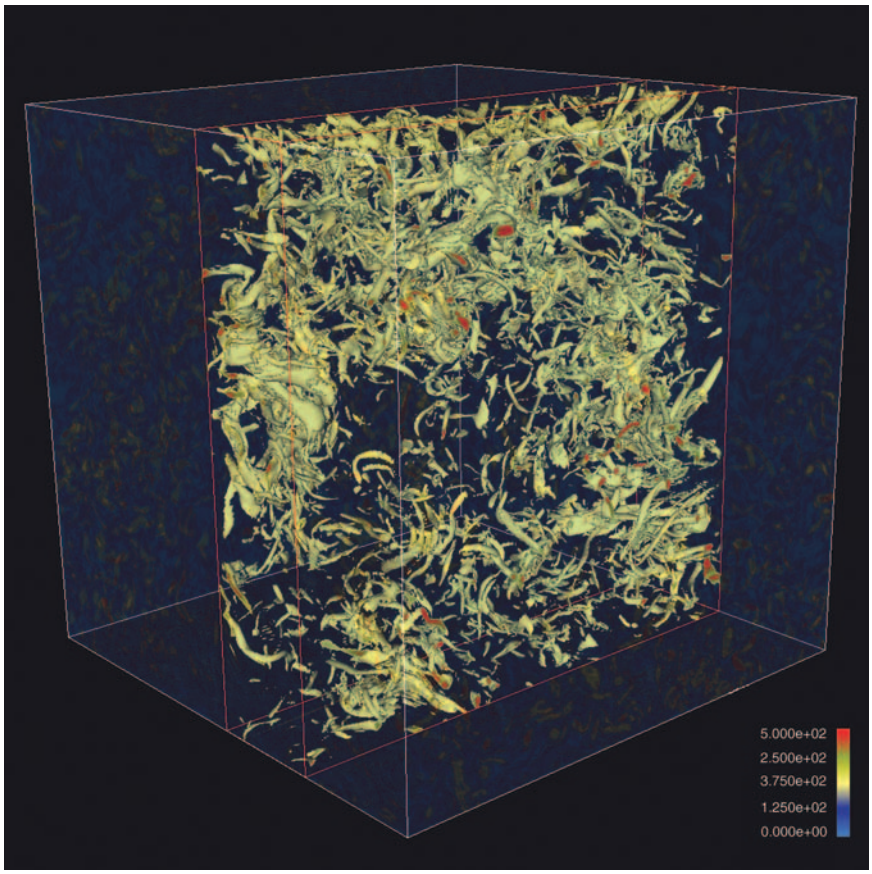


Figure 4. Surfaces of Constant Vorticity for Decaying Turbulence
This visualization is of the surfaces of constant vorticity magnitude in one of the 256^3 subdomains of the entire 2048^3 simulation (data set 5). There are 512 such subdomains in this simulation.

an underlying isotropic component (thick black line) that emerges from the angle-averaging procedure and seems to agree with the universally predicted two-fifteenths law to within 5 percent over a reasonable range of scales. This analysis (Kurien et al. 2004) reveals that the flux of helicity is more anisotropic and intermittent (in the sense of large departures from the mean) than the energy flux measured analogously by the four-fifths law (Figures 1 and 2).

In summary, angle averaging and statistical analysis have revealed that the isotropic component in turbulent flows is universal, agrees rather well with the Kolmogorov theory, and moreover, is consistent with the local version of Duchon and Robert (2000) and Eyink (2003). The procedure allows us to separate the contamination due to anisotropy from other effects, such as small Reynolds number and intermittency, that can muddy the measurement of clean scaling laws. The angle-averaging method also gives us a way to more efficiently use data and gain statistically significant results from single snapshots of the flow, whereas in the past, long time averages were taken, which led to data size and storage issues. Especially when we begin to start looking at the storage and analysis of data set 5, which needs of the order of 250 gigabytes of disk space, a scheme such as the angle-averaging procedure, which increases the amount of information we can glean from a single frame of turbulence data, is a definite asset.

Utility of Large-Scale Simulations

Our largest simulation (data set 5) is for a very highly resolved (2048^3), decaying flow at a moderate Reynolds number (270). The simulation's initial condition was taken from the centerline data gathered from a wind tunnel experiment performed at Johns Hopkins University (Tao et al. 2000). The simulation, performed on 2048 processors of ASC-Q, did not achieve the $R_\lambda \sim 700$ of the experiment. Therefore, direct comparison with the experimental results cannot be made until we can compute decaying flow at a higher Reynolds number or the experimental facility can rerun the experiment at a Reynolds number matching that of the existing simulation. However, a full numerical simulation provides access to the full spatial and temporal velocity field, while the experiments normally measure a sparse subset of the flow field.

Figure 4 shows the surfaces of constant vorticity magnitude for a 256^3 subdomain of the 2048^3 simulation. Vorticity visualizations are typically used to show the locations of the flow structures. In this case, the vorticity visualization shows that the generation of successively smaller energetic structures occurs by the stretching of regions of vorticity by the nonlinearity. The small structures in Figure 4 persist down to the grid size of the simulation.

Data sets 1 and 2 had Reynolds numbers similar to the number for data set 5 but only a quarter of the number of grid points to a side. That is, the linear size of the smallest scales resolved in the 512^3 simulations of data sets 1 and 2 was 64 times larger than the smallest scales in the 2048^3 simulation of data set 5. Because the 512^3 simulations cannot resolve almost two orders

of magnitude in scale that are accessible to the 2048^3 simulation, the coarser simulations obscure the turbulent fine structure seen at higher resolutions. Although they are quite suitable for observing the many inertial-range features described above, the coarser computations obscure the significant energetic events that occur at higher resolution. Clearly if we are to gain a deeper understanding of the spatial and temporal universal properties of turbulence through such numerical calculations, we must continue to pursue ways to compute larger resolved Navier-Stokes simulations and to develop efficient methods for analyzing the enormous quantities of data involved. ■

Further Reading

Duchon, J., and R. Robert. 2000. Inertial Energy Dissipation for Weak Solutions of Incompressible Euler and Navier-Stokes Equations. *Nonlinearity* **13**: 249.

Eliasen, E., B. Machenhauer, and E. Rasmussen. 1970. On a Numerical Method for Integration of the Hydrodynamical Equations with a Spectral Representation of the Horizontal Fields. In *Report No. 2*. Institute for Theoretical Meteorology, University of Copenhagen

Eyink, G. L. 2003. Local 4/5-Law and Energy Dissipation Anomaly in Turbulence. *Nonlinearity* **16**: 137.

Gotoh, T., D. Fukayama, and T. Nakano. 2002. Velocity Field Statistics in Homogenous Steady Turbulence Obtained Using a High-Resolution Direct Numerical Simulation. *Phys. Fluids* **14** (3): 1065.

Kang, H. S., S. Chester, and C. Meneveau. 2003. Decaying Turbulence in an Active-Grid-Generated Flow and Comparisons with Large-Eddy Simulation. *J. Fluid Mech.* **480**: 129.

Kolmogorov, A. N. 1941. The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Numbers. *Dok. Akad. Nauk. SSSR* **30**: 301.

Kurien, S. 2003. The Reflection-Antisymmetric Counterpart of the Kármán-Howarth Dynamical Equation. *Physica D* **175** (3–4): 167.

Kurien, S., and K. R. Sreenivasan. 2001. Measures of Anisotropy and the Universal Properties of Turbulence. In *New Trends in Turbulence: Turbulence Nouveaux Aspects: École de Physique DES Houches—Ujf and Inpg—Grenoble, a NATO Advanced Study Institute, Les Houches, Session LXXIV, 31 July–September 1, 2000*. p. 53. Edited by M. Lesieur, and F. David. New York: Springer-Verlag.

Kurien, S., M. A. Taylor, and T. Matsumoto. 2004. Isotropic Third-Order Statistics in Turbulence with Helicity: the 2/15-Law. *J. Fluid Mech.* **515**: 87.

Moin, P., and K. Mahesh. 1998. Direct Numerical Simulation: A Tool for Turbulence Research. 1998. *Annu. Rev. Fluid Mech.* **30**: 539.

Orszag, S. A. 1970. Transform Method for the Calculation of Vector-Coupled Sums: Application to the Spectral Form of the Vorticity Equation. *J. Atmos. Sci.* **27** (6): 890.

Overholt, M. R., and S. B. Pope. 1998. A Deterministic Forcing Scheme for Direct Numerical Simulations of Turbulence. *Comp. Fluids* **27** (1): 11.

Pope, S. B. 2000. *Turbulent Flows*. Cambridge, United Kingdom: Cambridge University Press.

Rogallo, R. S. 1981. Numerical Experiments in Homogeneous Turbulence. NASA Technical Report TM81315.

Sreenivasan, K. R., S. I. Vainshtein, R. Bhiladvala, I. San Gil, S. Chen, and N. Cao. 1996. Asymmetry of Velocity Increments in Fully Developed Turbulence and the Scaling of Low-Order Moments. *Phys. Rev. Lett.* **77** (8): 1488.

Sullivan, N. P., S. Mahalingam, and R. M. Kerr. 1994. Deterministic Forcing of Homogeneous, Isotropic Turbulence. *Phys. Fluids* **6** (4): 1612.

Tao, B., J. Katz, and C. Meneveau. 2000. Geometry and Scale Relationships in High Reynolds Number Turbulence Determined from Three-Dimensional Holographic Velocimetry. *Phys. Fluids* **12** (5): 941.

Taylor, M. A., S. Kurien, and G. L. Eyink. 2003. Recovering Isotropic Statistics in Turbulence Simulations: The Kolmogorov 4/5th Law. *Phys. Rev. E* **68** (2): 26310.

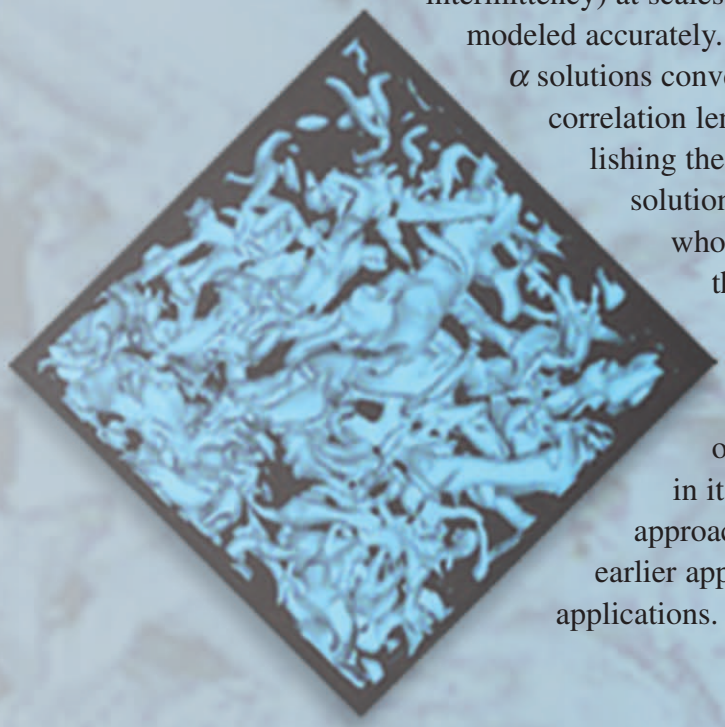
For further information, contact Susan Kurien (505) 665-0148 (skurien@lanl.gov) or Mark Taylor (505) 284-1874 (mataylo@sandia.gov).

The LANS-α Model for Computing Turbulence

Origins, Results, and Open Problems

Darryl D. Holm, Chris Jeffery, Susan Kurien, Daniel Livescu, Mark A. Taylor, and Beth A. Wingate

Over the last 50 years, numerous computational turbulence models have been proposed for obtaining closure. Obtaining closure means capturing the physical phenomenon of turbulence at computably low resolution, by mimicking the effects of the small scales on the larger ones without calculating them explicitly. The Lagrangian-Averaged Navier-Stokes alpha (LANS-α) model is the first to use Lagrangian averaging to address the turbulence closure problem. LANS-α modifies the nonlinearity of the Navier-Stokes equation, instead of its dissipation, thereby providing an alternative way to reach closure without enhancing viscosity. The LANS-α model arose from an educated guess, based on combining Lagrangian-averaged nonlinearity with Navier-Stokes viscosity. Its derivation from these first principles implied mathematical theorems for its solutions, thereby guaranteeing that the most basic properties of the flow (energy transport, circulation, variability, instability, dissipation anomaly, and intermittency) at scales above the effective cutoff scale of alpha are all modeled accurately. Mathematical analysis also proved that the LANS-α solutions converge to Navier-Stokes solutions in the limit as the correlation length parameter (alpha) tends to zero, thereby establishing the LANS-α model's accuracy. Moreover, the model's solutions for nonzero alpha possess a global attractor whose fractal dimension is finite, thus guaranteeing that the solutions are rigorously computable using finite resolution. The theorem-based approach of the LANS-α model has raised the mathematical standards for deriving other computational models of turbulence. Application of the alpha model is still in its infancy, but results so far suggest that this new approach will complement, and in some cases subsume, earlier approaches for modeling turbulence in real-world applications.



Turbulence is an outstanding unsolved multiscale nonlinear problem of classical physics. It occurs spontaneously in a fluid, when forcing by stirring at the large scales gets transferred by nonlinearity into slender, swirling circulations in the flow. These coherent swirling "blobs" of fluid, pierced by vortex lines and bounded by material circulation loops are called eddies. The eddies are Lagrangian structures, that is, they travel with the flow, stretching themselves into extended shapes (sheets or tubes) as they follow the flow induced by the vortex lines that pierce them. The coherent eddies, sheets, and tubes of vorticity, stretching themselves into finer and finer shapes, comprise the "sinews" of turbulence.

The characteristic features of turbulence—its distribution of eddy sizes, shapes, speeds, vorticity, circulation, nonlinear convection, and viscous dissipation—may all be captured by using the exact Navier-Stokes equations. The Navier-Stokes equations correctly predict how the cascade of turbulent kinetic energy and vorticity accelerates and how the sinews of turbulence stretch themselves into finer and finer scales, until their motions reach scales of only a few molecular mean free paths, where they may finally be dissipated by viscosity into heat. However, the fidelity of the Navier-Stokes equations in capturing the cascade of turbulence is also their downfall for direct numerical simulations of turbulence.

The number of active degrees of freedom required to simulate the turbulent cascade in high-Reynolds-number flows quickly outstrips the numerical resolution capabilities of even the largest computer. To make turbulence computable, scientists have developed various approximate models that halt the cascade into smaller, faster eddies. In most models, this effect is accomplished by causing the eddies below a certain size to dissipate computationally into heat. This dissipative imperative causes errors, however, because it damps out the variability in the larger-scale flow caused by the myriad of small scales of motion interacting nonlinearly together in the fields of the larger motion.

Opposite page: The sinews of turbulence are illustrated by level surfaces of vorticity calculated with the LANS-α model at a spatial resolution of 256^3.

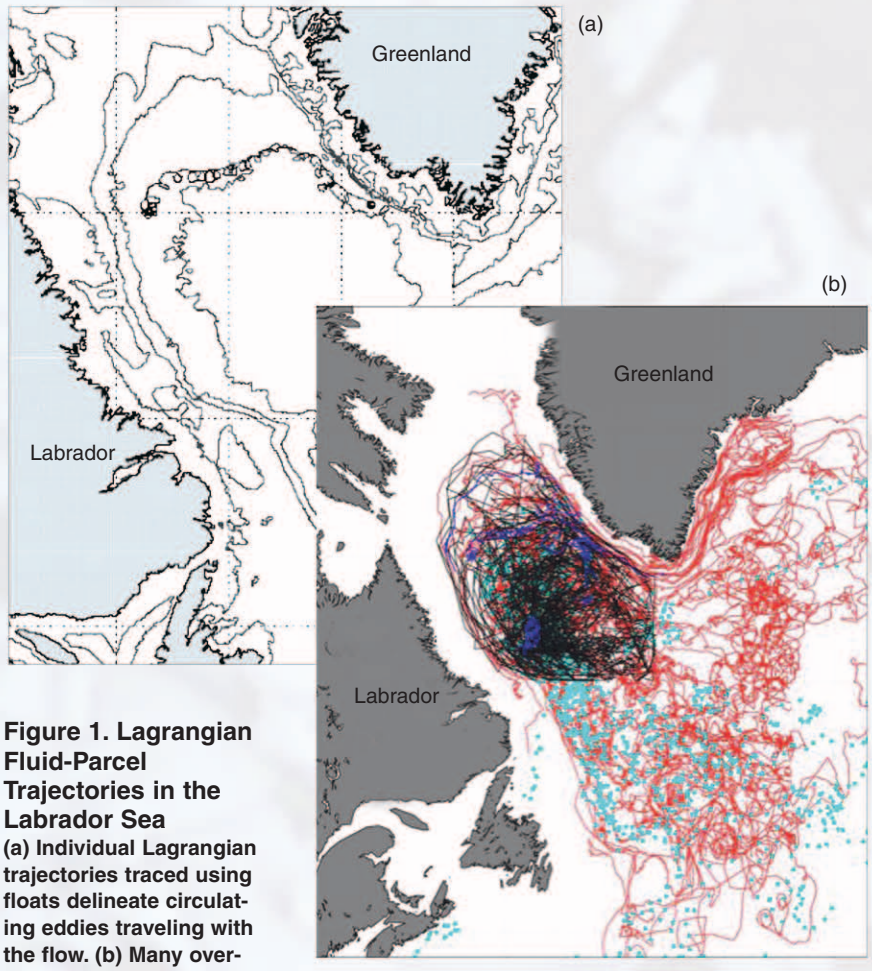
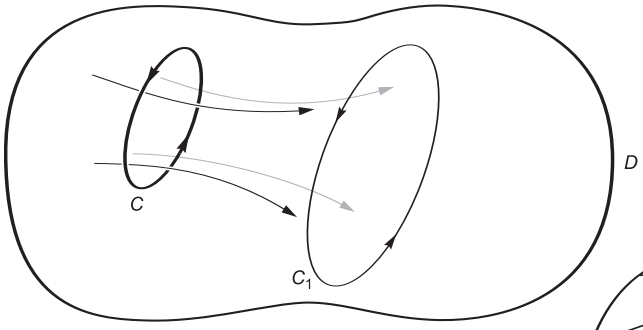


Figure 1. Lagrangian Fluid-Parcel Trajectories in the Labrador Sea
(a) Individual Lagrangian trajectories traced using floats delineate circulating eddies traveling with the flow. (b) Many overlapping trajectories capture the tangle of motions present in the flow. (Permission granted by Gerd Krahmann, Lamont-Doherty Earth Observatory of Columbia University.)

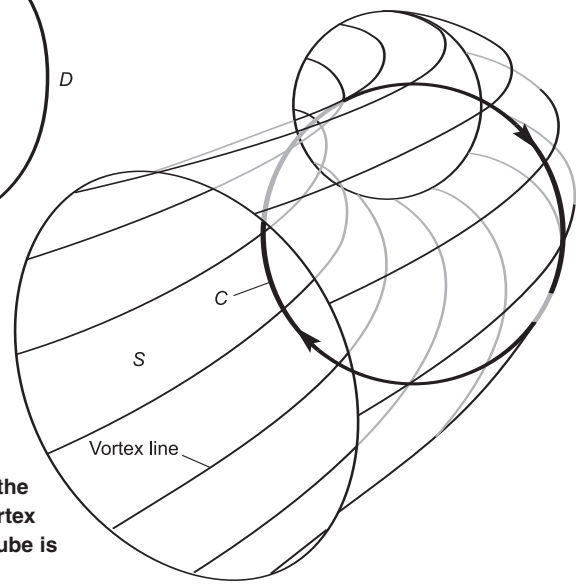
Consider the problem of modeling the average effects of turbulence on ocean currents in the North Atlantic Ocean. The North Atlantic contains circulations ranging in size from thousands of kilometers to only a few meters. The variability in the flow has been documented through observations of Lagrangian trajectories (trajectories moving with the fluid parcels) in the Labrador Sea. As shown in Figure 1 (Krahmann and Visbeck 2003), the Labrador Sea is full of highly oscillatory Lagrangian trajectories delineating the circulating eddy activity at the "mesoscale" size of tens of kilometers. Standard turbulence models for ocean simulations remove the fluctuating effects of all the scales of motion smaller than about 30 to 100 kilometers. Thus, the energy and information from the smaller scales are lost, and the resulting models ultimately are overdamped and inaccurate to the extent that the variability of their solutions depends upon these smaller scales.

The Lagrangian Eddy

A fluid possesses circulation if the integral of the tangential component of its velocity around any closed loop moving with the fluid is nonzero. A geometrical object such as a circulation loop embedded in, or traveling with, the fluid flow is an example of a Lagrangian quantity. A theorem of vector calculus by Kelvin and Stokes links the fluid's circulation with its vorticity, defined as the curl of its velocity. Namely, the circulation integral around the Lagrangian loop moving with the fluid is equal to the integral of the normal component of the fluid's vorticity, taken over any surface which has the circulation loop as its boundary. (This surface integral defines the "vorticity flux" through the surface whose boundary is the circulation loop.) Thus, circulation loops enclose distributions of vorticity flux, which may be regarded as bundles of vortex lines embedded in the fluid and wrapped by these Lagrangian circulation loops. These Lagrangian structures are known as "eddies." When the eddies stretch themselves into tubes, they are called "vortex tubes".



Above: As a material loop initially at *C* is carried by the fluid flow, it deforms to C_1 at a later time in domain *D*.



Right: A vortex tube is a material surface *S* surrounding a bundle of vortex lines (that is, lines tangent to the vorticity). The surface *S* is formed by a union of material loops *C*, each carried by the fluid flow. The divergence theorem implies that the flux of vorticity is the same through any slice, all along the vortex tube. Kelvin's theorem implies this flux of vorticity along the tube is constant in time. Thus, vortex tubes are "coherent structures."
(Redrawn from J. E. Marsden and T. S. Ratiu, *Geometric Analysis Methods in Fluid Mechanics*, manuscript in preparation.)

Capturing the mean effects of the smaller-scale circulations on the larger-scale motions in turbulence is called closure. In a novel approach, the Lagrangian-Averaged Navier-Stokes alpha (LANS-α) model we discuss here provides closure by modifying the nonlinearity in the Navier-Stokes equations to stop the cascading of turbulence at scales smaller than a certain length, but without introducing extra dissipation. Statistically, the size alpha in the LANS-α model is the typical distance that a Lagrangian trajectory fluctuates away from its time-mean trajectory. Hence, by definition, alpha is the smallest eddy scale still participating actively in the cascade. Eddies at scales smaller than alpha are, in effect, slaved to the mean motions of the larger ones; that is, they fluctuate locally as they are carried along in the frame of motion of the larger scales. This modification of the Navier-Stokes nonlinearity, derived by applying Lagrangian averaging techniques, allows the turbulence problem to remain computable at the resolution size of alpha, but to still retain the mean circulation effects of the smaller (subgrid) scales on the resolved solution. The LANS-α model is the first turbulence closure model to use Lagrangian averaging, from which it derives its name.

We shall briefly review the development of the LANS-α model from 1992 to1997, catalog its key results from 1997 to 2004, and finally discuss the open problems. The year 1997 was a turning point because only then was it realized that the ideas being developed in the context of ocean modeling had the potential to be used as a computable turbulence model.

The Development of the LANS-α Model

The origins of the LANS-α model can be traced to a one-dimensional model of nonlinear shallow-water wave dynamics, written down in a moment of inspiration on a blank page, in a pocket calendar, during a seminar in 1992 at the Center for Nonlinear Studies. Researchers began to take the equation seriously when it was discovered to be a soliton equation. That is, its initial value problem was found to possess exact nonlinear (weak) solutions, playfully dubbed "peakons" because of their sharp peaks, whose motion and interactions could be completely solved using elastic collision rules (Camassa and Holm, 1993). Subsequently, the equation was derived from Hamilton's principle of least action, which allowed it to be generalized to higher dimensions. The synergy between variational principles for soliton mathematics and dynamical concepts for turbulence modeling was developed further in the context of geophysical fluid dynamics, using a variety of approaches, including dominant asymptotics (Camassa et al. 1996, 1997).

The dominant asympotics technique produces hierarchies of equations that, at each increasing order in the asymptotic expansion, include more physics. Between 1993 and 1996, an interesting relation was discovered between standard dominant asymptotics and asymptotics performed on the Lagrangian in Hamilton's principle (HP). Namely, applying asymptotics in HP (before taking its variation) introduces terms in the resulting equations of motion that would ordinarily be dropped in dominant asymptotics, but which restore important fluid dynamical properties. These properties include conservation of both energy and potential vorticity (which arise from symmetries of the Lagrangian in HP) in the absence of viscosity, and preservation of Kelvin's theorem, which insures the proper nonlinear dynamics of circulation.

In 1996, Ivan Gjaja and Darryl Holm took the HP asymptotics idea a step further, while working on wave–mean flow interaction (WMFI) theory for ocean dynamics. WMFI theory addresses, for example, how surface waves can transfer momentum into regions far from their source. By applying Lagrangian averaging, as well as HP asymptotics, to a Wentzel-Kramer-Brillouin (WKB) wave packet representation of the rapid fluctuations, they derived the Gjaja-Holm WMFI equations, an asymptotic hierarchy of new equations for the wave–mean flow interaction. (Lagrangian averaging has a double meaning here because Gjaja and Holm averaged the Lagrangian in HP over the rapid phases of the WKB circulations at fixed Lagrangian coordinates.) Remarkably, these equations coincided with the result of applying dominant asymptotics and Lagrangian averaging to the exact Euler-Boussinesq equations for rotating, stratified, incompressible flows of an ideal fluid. This meant that the conservation laws for the Gjaja-Holm WMFI equations were programmed into the Lie-group symmetries of an averaged Lagrangian.

The Gjaja-Holm WMFI equations were developed in the context of the Laboratory's Climate Change Prediction Program, led by Robert Malone. They were intended to provide a turbulence model for rotating stratified fluids such as the oceans and the atmosphere. However, these WMFI equations were quite different from the usual turbulence models, and they needed to be simplified considerably before they could be recognized as a turbulence model. The inviscid part of the simplification was proposed in 1997, in work by Darryl Holm, Jerry Marsden, and Tudor Ratiu (1998a, 1998b). In this work, the Lagrangian-averaged Euler-alpha (LAE-α) equations, a Lagrangian-averaged closed form of the Euler equations (Navier-Stokes without viscous dissipation), were obtained. The key step in obtaining these LAE-α equations was the assumption of Taylor's "frozen-in" hypothesis, namely, that the mean statistics of the rapid fluctuations were carried along, or frozen, into the Lagrangian mean flow instead of propagating as wave packets, as had been assumed in deriving the Gjaja-Holm WMFI equations. Nonetheless, the parameter α^2 in the LAE-α equations has the same meaning as it does in the Gjaja-Holm WMFI equations. That is, α^2 is the typical size (statistical correlation length) of the excursions of a fluid parcel trajectory away from its mean (phase-averaged) trajectory, where the phase average is taken at a fixed Lagrangian coordinate along that trajectory. The derivation of the LAE-α equations using this form of Taylor's hypothesis is discussed in "Taylor's Hypothesis, Hamilton's Principle, and the LANS-α Model for Computing Turbulence" on page 172.

Once the LAE-α equations were derived, the stage was set for introducing viscosity and interpreting the resulting equations as a turbulence model. This last step in deriving the LANS-α model was taken in the collaboration among Shiyi Chen, Ciprian Foias, Darryl Holm, and Edriss Titi (1997–1998), when Foias, Titi, and their students Eric Olson and Shannon Wynne were visiting scholars at the Laboratory's Center for Nonlinear Studies (CNLS) and Institute for Space and Planetary Physics (IGPP). The introduction of viscosity was made first on an ad hoc basis, and then the LANS-α model was interpreted and confirmed as a turbulence model by comparing its predictions with experiment and

numerical simulations and by analyzing its theoretical properties.

How the LANS-α Model Differs from Others

As mentioned above, the key difference between the LANS-α model and other models of turbulence arises from the difference in the averaging technique used to derive the nondissipative LAE-α equations. In the LANS-α model, the average effects of the small scales on the large are modeled in the Lagrangian frame, which moves with the fluid parcels, instead of being modeled in the Eulerian frame, which is fixed in space. The Lagrangian averaging procedure leads to a new closure mechanism, a mechanism which reduces the number of degrees of freedom in the turbulence problem and approximates the effects of the small scales on the large. That new closure mechanism is based on nonlinear transport. In contrast, the more traditional Eulerian-averaging procedure leads to closure through linear or nonlinear diffusion.

Traditional Eulerian turbulence models use the Reynolds decomposition to separate the fluid velocity $\mathbf{u}$ at a point $\mathbf{x}$ into its mean and fluctuating components as $\mathbf{u} = \bar{\mathbf{u}} + \mathbf{u}'$, where $\bar{\mathbf{u}}' = 0$ and the overbar denotes an Eulerian mean (time average at a fixed point in space). Mathematically, Eulerian averaging commutes with the partial derivatives in space and time, but it does not commute with the advective, or material, time derivative $D/Dt = \partial/\partial t + \mathbf{u} \cdot \nabla$. This lack of commutivity between Eulerian averaging and the material time derivative leads to the unknown Reynolds stresses in the motion equations for the Eulerian mean velocity $\bar{\mathbf{u}}$ and, subsequently, to the well-known closure problem (see page 132 of the article "The Turbulence Problem"). In contrast, Lagrangian averaging commutes (by definition) with the material time derivative to produce the generalized Lagrangian mean (GLM) equations. These GLM equations, however, are also not yet closed. Moreover, Lagrangian averaging does not commute with spatial gradients. As a result, the Lagrangian-mean theory is history dependent, preserving the memory of its initial labeling along its Lagrangian trajectories, and the statistics of the Lagrangian-trajectory fluctuations must be prescribed in order to close the GLM equations.

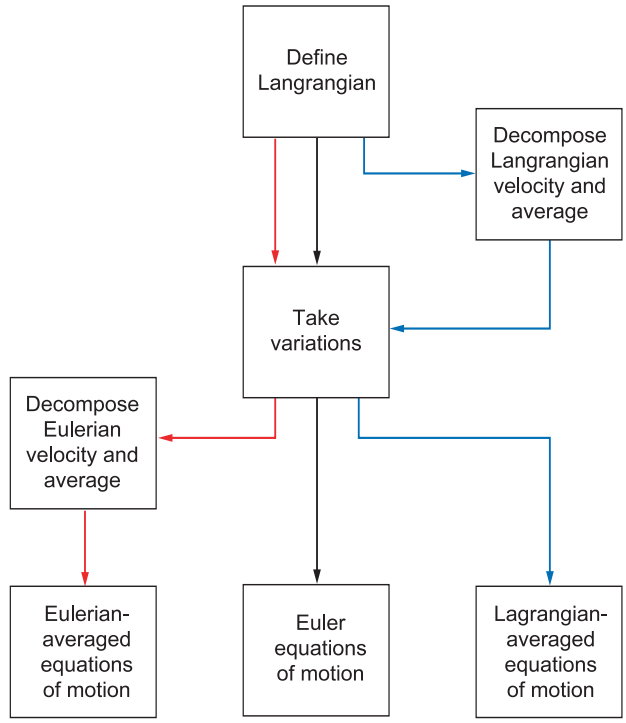


Figure 2. Paths to Derive Three Different Equations of Motion for Inviscid, Incompressible Fluid Flow
The blue path starts by decomposing the Lagrangian velocity into mean and fluctuating parts and then taking variations with respect to the Lagrangian averaged quantities to derive the LAE-α equations for ideal (inviscid) fluids.

Figure 2 illustrates the paths taken to derive three different sets of equations: the Euler equations for inviscid, incompressible flow (black), the corresponding Eulerian-averaged equations for the mean motion (red), and the inviscid LAE-α equations (blue). To produce the exact Euler equations of motion, first the Lagrangian in Hamilton's principle for fluids is defined and then the variations of the action (that is, the time integral of the Lagrangian) are taken. In turbulence models based on Eulerian averaging, most of the modeling effort takes place after Hamilton's principle of stationary variations of the action has produced the equations of motion. For Reynolds-averaged turbulence models, the velocity is then decomposed into its (Eulerian) mean and fluctuating quantities, or for the large eddy simulation (LES) framework, the equations in the Eulerian frame are spatially filtered. In contrast, for the LAE-α framework, the modeling occurs in averaging the Lagrangian in Hamilton's principle before the variations are taken, and the Lagrangian-averaged equations result from taking variations of Lagrangian-averaged quantities using the Euler-Poincaré theory of Holm et al. (1998a, 1998b). (The averaged Lagrangian approach is

The LANS-α Model Equations

The LAE-α equations are

$$\frac{\partial \mathbf{v}}{\partial t} + \underbrace{\mathbf{u}\cdot\nabla\mathbf{v} + \nabla\mathbf{u}^{\mathrm{T}}\cdot\mathbf{v}}_{\text{Modified nonlinearity}} + \nabla p = 0 \ , \tag{1}$$

$$\text{with} \quad \nabla\cdot\mathbf{u} = 0, \quad \text{and} \quad \mathbf{v} = \left(1-\alpha^2\Delta\right)\mathbf{u} \ . \tag{2}$$

Rewriting Equation (1) as the time rate of change of momentum in the frame of the moving fluid yields

$$\frac{d}{dt}\left[\mathbf{v}\left(t,\mathbf{x}(t)\right)\cdot d\mathbf{x}(t)\right] + \nabla p\cdot d\mathbf{x}(t) = 0 \quad \text{along} \quad \frac{d\mathbf{x}}{dt} = \mathbf{u}\left(t,\mathbf{x}(t)\right) \ . \tag{3}$$

Adding viscosity and forcing yields the LANS-α equations:

$$\frac{\partial \mathbf{v}}{\partial t} + \underbrace{\mathbf{u}\cdot\nabla\mathbf{v} + \nabla\mathbf{u}^{\mathrm{T}}\cdot\mathbf{v}}_{\text{Modified nonlinearity}} + \nabla p = \underbrace{\nu\Delta\mathbf{v} + \mathbf{f}}_{\text{Viscosity \& forcing}} \ , \tag{4}$$

$$\text{with} \quad \nabla\cdot\mathbf{u} = 0, \quad \text{and} \quad \mathbf{v} = \left(1-\alpha^2\Delta\right)\mathbf{u} \ . \tag{5}$$

The Kelvin circulation theorem for the LANS-α model is

$$\frac{d}{dt}\oint_{c(\mathbf{u})} \mathbf{v}\cdot d\mathbf{x} = \oint_{c(\mathbf{u})} \left(\nu\Delta\mathbf{v} + \mathbf{f}\right)\cdot d\mathbf{x} \ . \tag{6}$$

much simpler and more transparent than averaging the equations term by term, and a theorem guarantees that the same equations result in either order. A concise description of this process is given in the article "Taylor's Hypothesis, Hamilton's Principle, and the LANS-α Model for Computing Turbulence"on page 172.) The LAE-α equations (in terms of Eulerian averaged quantities) are given by Equations (1) and (2) in the box above.

The two velocities **u** and **v** in the LAE-α Equations (1) and (2) are averaged quantities. However, the transport velocity **u** is smoother than the transported velocity **v** by inversion of the Helmholtz operator, $(1 - \alpha^2\Delta)$. This inversion operation amounts to obtaining velocity **u** by filtering velocity **v** over the length scale α. When $\alpha \to 0$, then $\mathbf{v} \to \mathbf{u}$, and one recovers the original Euler equations.

According to the Euler-Poincaré theory of Holm et al. (1998a, 1998b), the transport velocity **u** in Equation (1) is the average velocity at which the fluid material moves. So, what is the interpretation of the other average velocity **v** in Equation (2)? The Euler-Poincaré theory defines the velocity **v** as the momentum per unit mass of the Lagrangian averaged motion. This momentum is obtained by taking the variational derivative of the averaged Lagrangian in Hamilton's principle with respect to the average velocity **u**. The two velocities differ for the usual reason, namely, that nonlinearity and averaging do not commute. One may understand the different roles of these two velocities by considering the LAE-α equation as a form of Newton's law for the time rate of change of the momentum in the frame of fluid motion. Namely, Equation (1) is equivalent to Equation (3). Thus, the second term in the modified nonlinearity of Equation (1) arises from the rate of change of the line element $d\mathbf{x}(t)$ in the frame of motion of the fluid moving with velocity **u**. (Of course, the first term in this nonlinearity arises from the chain rule.)

After deriving these inviscid LAE-α equations, we added viscosity and forcing so that energy would decay and momentum would diffuse, thereby obtaining the LANS-α model Equations (4) and (5). When $\alpha \to 0$, then $\mathbf{v} \to \mathbf{u}$ and the LANS-α equations revert to the original Navier-Stokes equations.

Remarkably, the LANS-α equations answered an outstanding mathematical question going back to the early efforts of Leray (1934) to regularize the Navier-Stokes equations. This question was emphasized by Galovotti (1993), namely, "How does one regularize the Navier-Stokes equations without destroying their circulation properties?" (Recall that the LANS-α model was developed to deal with average effects of turbulence in ocean circulation.) The answer is obtained by direct calculation, which yields the Kelvin circulation theorem for the LANS-α equations given by Equation (6). Physically, this theorem means the circulation of the velocity **v** around a material loop c moving with smoothed transport velocity **u** is created by the integral around this loop c of the sum of the viscous and external forces. When $\alpha \to 0$, then $\mathbf{v} \to \mathbf{u}$, and one recovers the fundamental Kelvin circulation theorem for the Navier-Stokes equations, thereby regaining

the picture of the sinews of turbulence described earlier. The Kelvin circulation theorem for the LANS-α equations above shows how this picture is modified by Lagrangian averaging. We will discuss later how the LANS-α Equations (4) and (5) regularize the Navier-Stokes equations in the sense discussed by Leray (1934) and Galovotti (1993).

Results from 1997 to 2004

In the next few sections, we present a sampling of key results for the LANS-α model from 1997 to 2004. This is not meant to be an exhaustive review of the entire body of the LANS-α literature, but a sampling of theoretical and numerical results to give the reader a flavor for what is known and what remains to be studied.

Through the rest of this article, the word 'modeling' refers to the mathematical description of unknown quantities in terms of known quantities for the purpose of regularizing or reducing the number of active degrees of freedom in the Navier-Stokes equations.

"Benchmark" Tests of the LANS-α Model

Once we recognized that LANS-α might be interpreted as a turbulence model, we tested this hypothesis by using LANS-α to calculate some of the classic turbulence problems. These included turbulent flow in a pipe, forced turbulence in a periodic domain, and decay of turbulence in a periodic domain. In all three cases, the results were very encouraging.

LANS-α Stationary Solutions for Pipe Flow Compared with Experimental Data. Figure 3 (Chen et al. 1999a) shows a semilog plot of the time-averaged velocity for turbulent flow in a pipe vs distance from the wall at three different Reynolds numbers. The experimental data (solid lines) were measured at the Princeton "super pipe" and correspond to turbulent flows with the highest values of Reynolds number available in a pipe-flow experiment (Zagarola 1996). The dashed lines show the corresponding stationary solutions of the LANS-α model. All three solutions were obtained using a single constant value of alpha (equal to about one percent of the pipe radius). This value of alpha was obtained by matching the first set of data at a Reynolds number of 98,812. Then, alpha was held constant for the other two comparisons. The family of mean velocity profiles $\phi(\eta)$ seems to possess a lower envelope. This straight line in the semilog plot satisfies the famous von Kármán logarithmic law of the wall. However, the LANS-α steady solutions match the experimental data all the way across the pipe flow domain, from a few tens of wall units away from the pipe boundary all the way to the pipe center, where the peak of each curve occurs. (These peaks are offset because the wall unit η contains the Reynolds number in its definition.)

Note that the LANS-α solution matches the measured mean velocity over many orders of magnitude in wall units. That agreement is a good sign because turbulence models must describe a wide range of scales of motion—from the scale of the forcing down to the dissipation scale. The faint, dotted lines show the recent power law from Barenblatt-Chorin (1997), which does not capture the peaks of the curves. The excellent agreement with the experimental mean velocity profiles (from Chen et al. 1998, 1999a) provided the first clue that the LANS-α equations for the Lagrangian mean velocity might be interpretable as a model of turbulence.

Navier-Stokes Equations: Forced Turbulence in a Periodic Domain. Next, we tested the LANS-α model on the problem of forced turbulence in a three-dimensional (3-D) periodic domain where turbulence is approximately homogeneous and isotropic so that Kolmogorov-like scaling laws should obtain. We performed direct numerical simulations of the LANS-α model and examined the effect of increasing alpha on the energy spectrum $E(k)$, where k is the wave number. Results from Chen et al. (1999) show that, in the spectral region $k\alpha < 1$ (that is, for spatial scales larger than alpha), $E(k)$ is proportional to $k^{-5/3}$, as expected for homogenous, isotropic turbulence. In other words, the energy spectrum at these spatial scales is essentially unaffected by the presence of the α-modification (regularization). However, in the spectral region with $k\alpha > 1$ (that is, for spatial scales smaller than alpha), $E(k)$ rolls off faster as wave number increases. In Chen et al. (1999b), we kept alpha fixed at $\alpha = 1/8$ of the domain size and compared the energy spectrum for a high-resolution mesh of 256^3 cells and a low-resolution mesh of 64^3 cells. The energy spectra at the large scales (in the inertial range) were the same for both simulations, which means that, for this problem of forced turbulence, the large-scale flow properties can be preserved when the resolution is decreased by a factor of 8. (The actual computational savings is a factor of about $4^4 = 256$ in computer time.) This result implies that direct numerical simulation of the LANS-α model allows a significant computational savings over the direct numerical simulation of the Navier-Stokes equations.

Later, Foias et al. (2001) used dimensional arguments to predict the faster energy-spectrum rolloff for $k\alpha > 1$ that was seen in the computations. These dimensional arguments predict-

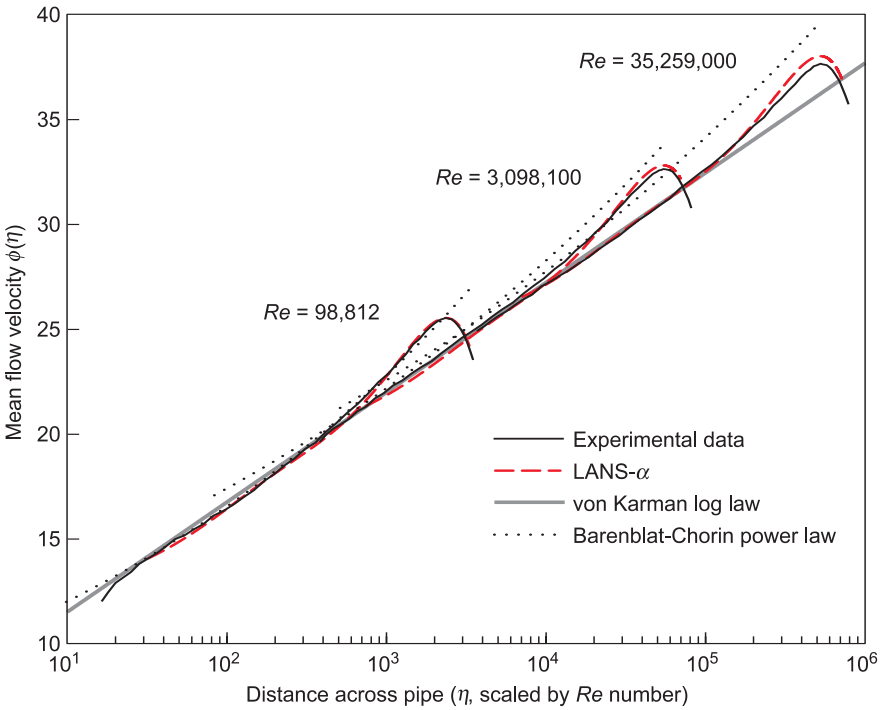


Figure 3. Mean Velocity Profiles for Pipe Flow
Comparison in this figure from Chen et al. (1998, 1999a) of mean flow profiles for turbulent flow in pipes given by experimental data shows reasonable agreement with the profile of the corresponding solution of the LANS-α equations at the highest experimentally available Reynolds numbers. Here, the mean-velocity profile in the pipe for the LANS-α equation (the red dashed line) is compared with the experimental data (the solid line) of Zagarola (1996). (Copyright 1998 by the American Physical Society.)

ed the rolloff to be $k^{-5/3} \rightarrow k^{-3}$ for the LANS-α model. The rolloff $k^{-5/3} \rightarrow k^{-3}$ is consistent with the Re^2 scaling behavior in computational work for a fully resolved direct numerical simulation of the LANS-α equations, in comparison with the Re^3 scaling behavior in computational work for a fully resolved direct numerical simulation of the Navier-Stokes equations.

The relative scaling of Re^2 for LANS-α vs Re^3 for the Navier-Stokes equations implies a two-thirds-power scaling in the computational work required in the direct numerical simulation of the Lagrangian-averaged LANS-α equations vs the exact Navier-Stokes equations, provided the k^{-3} inertial range for the LANS-α model is resolved. At a large Reynolds number, *Re*, this scaling can provide a substantial savings in computational work.

Navier-Stokes Equations: Turbulence Decay in Three-Dimensions. A more stringent test of the LANS-α model is the initial value problem for 3-D incompressible turbulence known as turbulence decay. In this problem, one starts from a turbulent initial condition that results from forcing, and then one turns off the forcing and lets the turbulence decay away. In recent computations (Holm and Kerr 2002; Geurts and Holm 2002a, 2002b; Mohseni et al. 2000, 2001), numerical comparisons between large-eddy simulation (LES) methods and the LANS-α model were made for the onset, development, and decay of shear turbulence. All three of these numerical studies compared the predictions of the LANS-α model for the case of shear turbulence decay in three dimensions against the most advanced LES models, which achieve closure through modifying the energy diffusion rather than the nonlinearity. The standard of comparison for these low-resolution model simulations using the LANS-α model and several standard LES approximate models was a direct numerical simulation of the full Navier-Stokes equations at a much higher resolution. In these investigations, Holm and Kerr started from a Taylor-Green initial condition specified by spectral data; Geurts and Holm started from the classic physical realization of the Kelvin-Helmhotz instability, leading to the formation and decay of turbulent shear layers; and Mohseni et al. studied the decay of turbulence in the standard Comte-Bellot and Corrsin wind-tunnel configuration.

In all these benchmark problems, the results of the Lagrangian-averaging approach to modeling turbulence were found to be comparable with the best of the standard LES approximate models.

Relation of LANS-α Model to Large-Eddy Simulations

LES models are often used in numerical simulations of turbulence. Because of their importance and their formal similarity to LANS-α, considerable work has been devoted to understanding the connection between LANS-α and LES.

The basis for the LES approach is spatial filtering of the Navier-Stokes equations in the Eulerian frame, whereas the theoretical basis for obtaining the LANS-α equations is Lagrangian averaging. Of course, both approaches face difficulties with closure. Either approach to closure introduces approximations because the equations are nonlinear, and neither the averaged nor the filtered product of two factors would be equal, in general, to the product of the averaged, or filtered, factors.

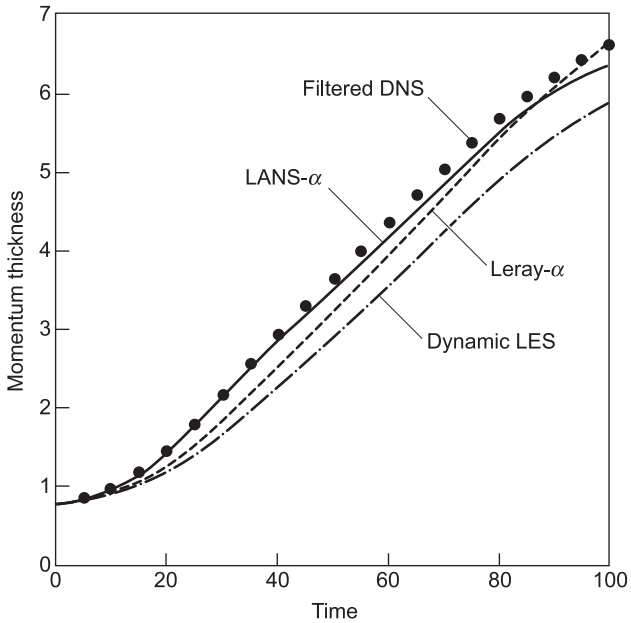


Figure 4. Comparing Results of LANS-α and Leray-α with Dynamic LES for the Turbulent Mixing Layer

This figure from Geurts and Holm 2002 compares the momentum thickness as a function of time for the turbulent mixing layer initiated by the Kelvin-Helmholtz instability. Here $\alpha = L/16$ and three LES models are plotted: LANS-α (solid), Leray-α (dashed), Dynamic LES (dash-dotted). The nearly grid independent DNS solution at resolution is shown as solid circles. The momentum thickness for the mixing layer begins with a strong convective surge, which the LANS-α model follows well. The Leray-α model lacks the term that provides line-element stretching to complete Kelvin's circulation theorem, and apparently this term is important at an early time. Dynamic LES apparently lags in the beginning and never catches up, perhaps because it attempts to model nonlinear turbulent transport as diffusion.

(Reprinted with the permission of Springer-Verlag.)

Formally, the Lagrangian-averaged turbulence equations appear similar to the LES turbulence equations (Domaradzki and Holm 2001), but there are significant differences in the interpretations of their solutions. These differences in interpretation arise because the two models are derived from different fundamental principles. The similarity between them arises because both approaches yield expressions for conservation, or balance, of momentum. The similarity between them also arises through interpreting the equations produced by the Lagrangian-averaging approach as embodying a "regularization principle," which involves an explicit filter and its inversion (Guerts and Holm 2003). Momentum conservation for this regularization principle identifies the stress tensor corresponding to the implied subgrid model, which resolves the closure problem. Thus, the model equations resulting from the Lagrangian-averaged turbulence method convey a central and very specific physical role to a filter: The transport velocity is a filtered version of the fluid momentum, including the mean momentum of the fluctuations. This role differs from that of the filter in the foundations of the LES approach. In the LES approach, the difference between the filtered product of velocities and the product of filtered velocities is modeled as a symmetric tensor involving gradients of the filtered velocity, whose divergence introduces dissipation of energy.

In terms of physical effects, the dissipation introduced by LES filtering smoothes and slows the fluid's momentum, so the LES results tend to be sluggish compared to DNS and, thus, LES often fails to capture the true variability of turbulence. In contrast, the modification of the nonlinearity in the alpha model "enslaves" the smaller scales to the larger ones, and their circulation is not lost to heat. This feature gives the LANS-α model an advantage. For example, it produces sharper, more-pronounced coherent structures and higher variability than even the best LES models (the dynamic LES models) in computing turbulent shear mixing (see Figure 4).

Application of LANS-α to Specialized Problems

Thin-Layer Navier-Stokes Equations: Self-Similarity. Steady self-similar solutions (for the dependence of mean downstream velocity U in a two-dimensional (2-D) boundary layer of the form $U(x, y) = g(x)f(y/x)$) of the thin-layer Navier-Stokes (TLNS) equations were known for laminar boundary-layer problems since Paul Blasius in 1908. For turbulent shear flows such as jets, wakes, and plumes, the Kelvin-Helmholz instability generates mixing near the interface between the moving and stationary fluids, and the mixing region spreads tranversely, as the unstable entrainment interaction between the fluids proceeds in time—see Figure 5(a). Finding solutions to these self-similar flows was plagued by closure problems until Ludwig Prandtl (1925) invented the mixing-length theory, which captures the drag effects of turbulent eddies. Prandtl's mixing length is a macroscopic length scale defining the mean distance between eddy collisions; it was meant to be analogous to the mean free path

between molecules in kinetic theory. For most TLNS self-similar problems, such as jets, wakes, and plumes, analytical results from Prandtl's mixing length theory match experimental data reasonably well.

Except for that simple mixing-length theory, self-similar solutions of most turbulence models have not been investigated. However, because the LANS-α equations were derived to have self-consistent dynamics, such solutions seemed possible. Indeed, thin-layer self-similar solutions of the LANS-α model were found for boundary layers, jets, wakes, and plumes (Cheskidov 2002, Holm et al. 2003, Putkaradze and Weidman 2003). These solutions arise by introducing both α (a statistical property of Lagrangian averaging) and a mixing length (a statistical property of Eulerian averaging). Each averaging mechanism seems to control a different aspect of the analytical self-similar solutions. For example, in the planar jet shown in Figure 5(a), the thickness of the jet $g(x)$ depends only on x, the distance downstream from the source, and that thickness is determined entirely by mixing-length theory. On the other hand, the profile of the analytical solution for the mean velocity U across the jet—see Figure 5(b)—is a function of the similarity variable, $\eta = y/x$, and the shape of that profile is determined by α in these calculations. Figure 5(b) also shows a comparison of the alpha model's similarity solution with the experiments of Effie Gutmark and Israel Wygnanski (1976).

Understanding the interplay of diffusion (as in the mixing-length theory) and transport (as in the LANS-α model) is still an outstanding problem in modeling these self-similar turbulent flows.

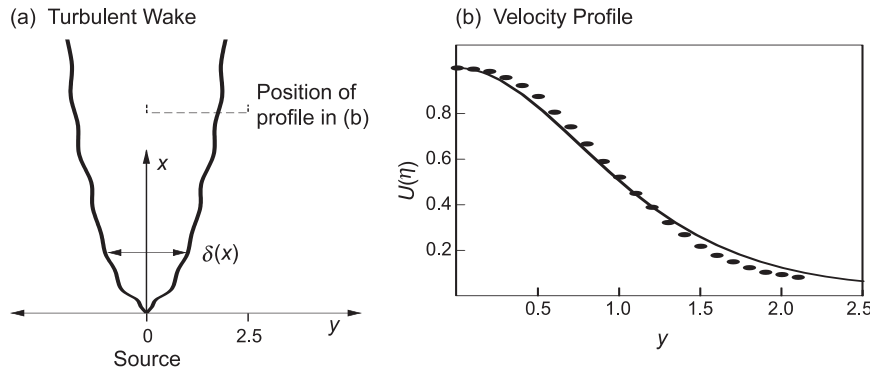


Figure 5. Self-Similar Solutions for a Planar Turbulent Jet
(a) A turbulent jet gushes out from a source. (b) The analytical solution of the LANS-α equations for a planar turbulent jet is compared with results from experiments) by Gutmark and Wygnanski (1976).
(Reprinted with the permission of Cambridge University Press.)

Geophysical Fluids. Geophysical fluid dynamics offers a unique regime in which to compare the LANS-α model with other well-known models because the energy does not cascade to the small scales as it does in 3-D incompressible Navier-Stokes turbulence. Instead, these quasi-2-D flows are characterized by an upscale transfer of energy to lower wave numbers. This transfer of energy creates the large-scale vortices observed in nature. As a consequence, coarse-resolution models have a good chance of simulating the most important dynamical features of these flows. Between 1997 and 2004, two important regimes were studied: quasi-geostrophy, whose principal wave solutions are slow-time-scale Rossby waves, and the rotating shallow-water equations, whose solutions include both Rossby waves and fast inertial waves.

Quasi-Geostrophy (QG). Application of the LANS-α model to slow, large-scale motions for rotating, planetary-scale fluid dynamics has yielded mixed results. Two sets of simulations have been performed of the problem of wind-forced circulation in a closed ocean basin.Wind-forced circulation results, ostensibly, in two counter-circulating gyres. As described in Greatbatch and Nadiga (2000), the time-mean ocean basin circulation predicted by the QG equations shows a four-gyre pattern, although its instantaneous motion generally shows only two gyres, which fluctuate strongly and rapidly.

In the low-resolution LANS-α simulations of Nadiga and Margolin (2001), the four-gyre time-mean pattern was recovered, but only after an appropriate combination of alpha and dissipation parameters were determined from a higher-resolution eddy-resolving run (regarded as a direct numerical simulation). Further, the correspondence between the time mean of the eddy-resolving run and the α-parameterized run was less than satisfactory and not fully understood. This incompleteness left open questions that still need further study.

In Holm and Nadiga (2003), an LES viewpoint was adopted, in which low-resolution simulations of the QG-α model and some of its close variants were compared with time means of direct numerical simulations of QG for the full double-gyre problem. This approach led to significantly improved results for the time-mean circulation in the double-gyre problem, and it also captured reasonable variability in the form of eddy kinetic energy and eddy potential enstrophy. Figure 6 shows contour plots of the time-averaged stream function, in

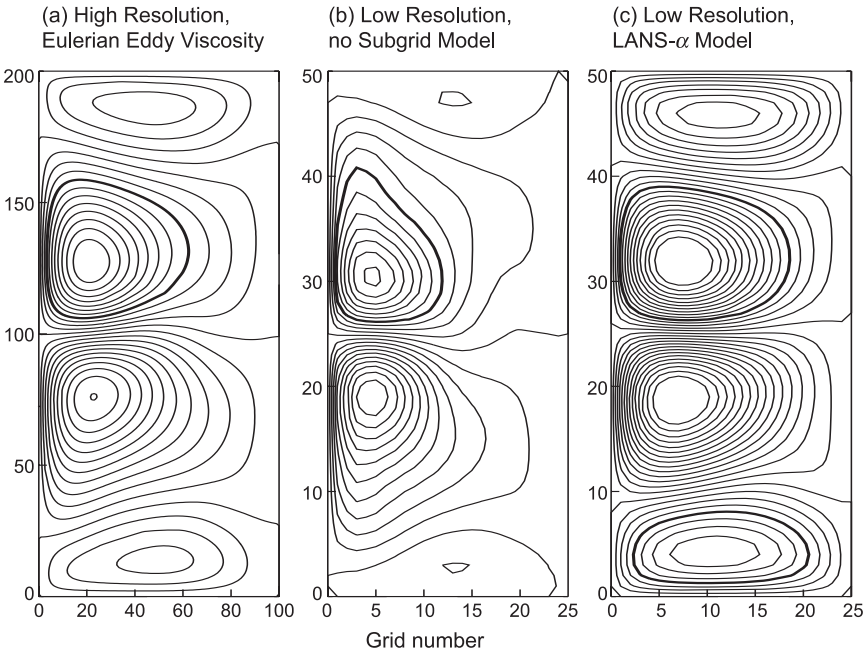


Figure 6. Quasi-Geostrophic Double-Gyre Problem
This figure from Holm and Nadiga (2003) shows time-averaged contour plots of the stream function for the quasi-geostrophic double-gyre problem. (a) Shown here is a 96^3 high-resolution QG simulation. The Munk layer scale is 0.02*L*, and the grid resolution is 0.01*L*. At this low level of viscosity, the time-mean stream function displays a four-gyre structure even though the wind forcing is that for a double gyre. (b) This simulation is run at a resolution that is 4 times coarser—a grid resolution of 0.04*L*. With no modeling of the subgrid scales, we find that the outer pair of gyres is greatly weakened compared with the pair in (a). (c) This simulation is also run at a resolution that is 4 times coarser, but it uses the alpha model to account for subgrid scale activity. Here, we find that the outer pair of gyres is restored. However, the strength of the wind-driven and the eddy-driven mean circulation is slightly higher than the resolved simulation shown in (a). We are currently studying the reasons for this overprediction. (Permission granted by the American Meteorological Society.)

which the four-gyre pattern clearly emerges.

The results in Figure 6 show that the LANS-α model yields a decided benefit in predicting the correct time-mean variability for this problem. However, the strength of the circulation was slightly higher than in the resolved simulation.

Rotating Shallow Water (RSW). Because RSW produces fast waves, the RSW equations are hard to solve numerically. The maximum allowable time step is $\Delta t \leq C/N$, where C is a constant of order unity and N is the number of mesh points in the domain. Using the LANS-α model to simulate these equations led to a slowing down of the fastest waves (those with wave lengths smaller than α). Consequently, LANS-α simulations that used a much larger time step, given by $\Delta t^{(\alpha)} \leq C\alpha$, retained the high variability found in the highest-resolution runs. This means that refining the mesh with a fixed α causes the LANS-α model's maximum allowable time step to go to a constant, while the shallow-water model requires its time step to go to zero.

These simulations also revealed that the LANS-α model preserves the time variability of the dynamics. Figure 7 from Wingate (2004) shows the time series spectra for the kinetic and potential energy on two different grids for the double-gyre problem (see the caption for details).

Although the LANS-α model does reproduce the time variability of shallow-water flow, these results raise several questions. As shown in Figure 7, increasing α may cause an overprediction of variability, as discovered in the study of the double gyre in Nadiga and Holm (2003). This overprediction of variability leads us to ask, "How does one make an optimal choice of α?" Also, for the same viscosity, the alpha model typically has a higher variability than coarse-resolution simulations of the exact equations. This increased variability occurs because, in the LANS-α model, the enstrophy-like energy (not the translational kinetic energy) controls dissipation at high wave numbers. This result brings up the question, "Should the Reynolds number be defined differently in these cases?" This issue will be addressed in the section on open problems.

Modeling Fluid Instability

The stability and instability of flows in different parameter regimes (such as Reynolds number *Re*, Rossby number *Ro*, and Froude number *Fr*) could be altered, in principle, by introducing turbulence models. We performed two studies of fluid instability in the LANS-α model.

Elliptical Instability. The elliptical instability converts 2-D fluid motions into 3-D convection, so it provides a fundamental mechanism at the onset of turbulence. Motivated by the idea that a turbulence simulation method should not erroneously predict stability in a flow that is actually unstable, Fabijonas and Holm (2003) investigated the elliptical instability in the

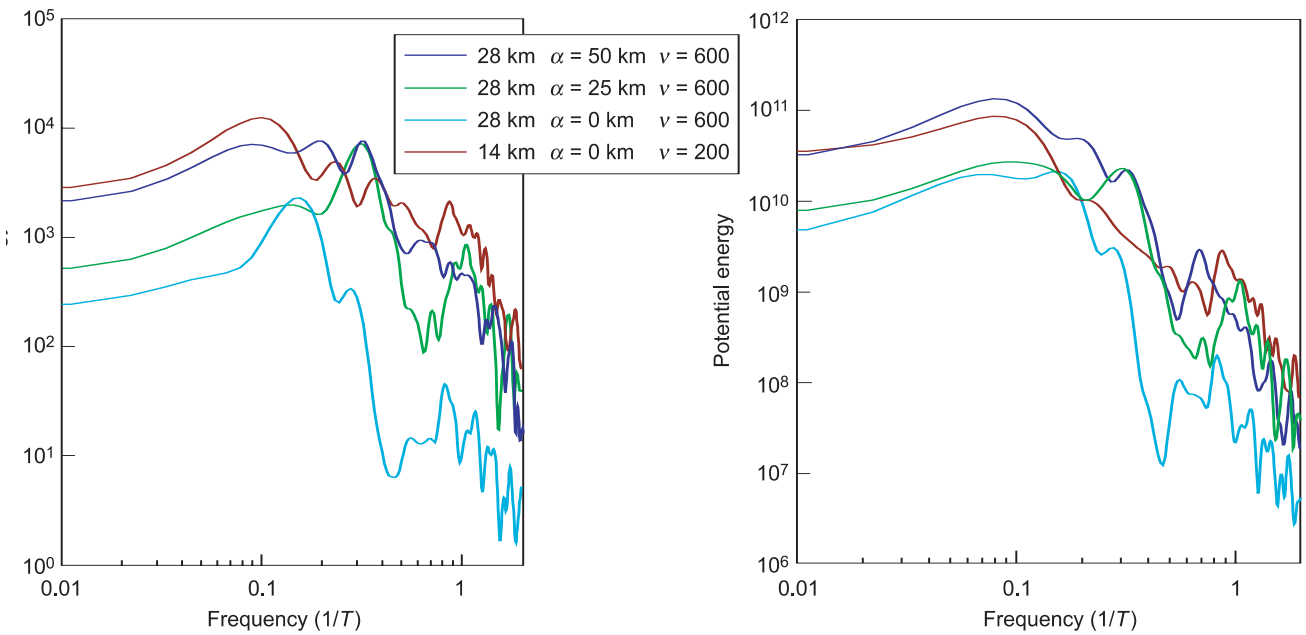


Figure 7. Improved Variability for LANS-α Shallow Water Simulations
(a) The kinetic energy is shown as a function of frequency; (b) the potential energy is shown as a function of frequency. In both (a) and (b), the values are for shallow water simulations at different values of α and different resolutions. The Rossby deformation radius for all cases is approximately 48 km. The high-resolution calculations (red) with $\alpha = 0$, an average grid spacing of 14 km that resolves the Rossyby radius, and a viscosity of 200 m^2/s serve as our standard of energy variability. The other three simulations were performed on a much coarser mesh with an average grid spacing of 28 km, a mesh size for which the Rossby deformation radius is not well resolved. The pale blue curve shows the results of the simplest eddy viscosity model ($\alpha = 0$ and just enough viscosity is added to prevent numerical instability). The flow is sluggish with almost an order of magnitude decrease in the variability of the kinetic energy due to the increase in the dissipation of the total energy. The purple curve shows the increased variability that results by introducing alpha at the value $\alpha = 25$ km. The dark blue curve shows that, by increasing alpha to the size of the Rossby deformation radius, we recover the variability of the fine-grid case.

LANS-α model and showed that the model preserves, but modifies, this important instability. In particular, the LANS-α model reduces the maximum growth rate for higher wave numbers, $k\alpha >> 1$, but for slightly lower wave numbers, $k\alpha > 1$, the model increases the maximum growth rate. This enhancement allows the dynamics of the small scales to affect the larger scales. This work led to a sequence of investigations: from early assessments of the average effects of turbulence on elliptic instability to later assessments of the combined effects and interplay of turbulence, rotation, and stratification on elliptical instablility.

Baroclinic Instability.
Investigations using global ocean models or coupled ocean, atmosphere, and ice models require the use of coarse meshes. The meshes are often so coarse that the Rossby deformation radius is not resolved,[1] and consequently baroclinic instability is incorrectly predicted. Baroclinic instability is initiated by vertical shear in a rotating, stratified flow and describes the process of converting available potential energy to kinetic energy on scales of the Rossby deformation radius. This is one of the most important dynamical phenomena in geofluid dynamics and one that any turbulence model must reproduce if it is to simulate the correct variability.

[1] The Rossby deformation radius is the distance at which the pressure force balances with the Coriolis force in the motion equation.

In Holm and Wingate (2004), neutral curves for the onset of baroclinic instability from the simplest LANS-α model were compared with those from the simplest eddy-viscosity model (see Figure 8). Neutral curves show the shear forcing required to initiate baroclinic instability versus wave number. Figure 8(a) presents LANS-α neutral curves for three values of αk_{int}, the length of α relative to the Rossby deformation radius. As α, or αk_{int}, is increased, the critical wave number (wave number at the minimum of the neutral curve) shifts to lower wave number while the value of the minimum forcing required remains the same. Thus, the onset of baroclinic instability remains resolvable with fewer grid points.
Figure 8(b) shows neutral curves for

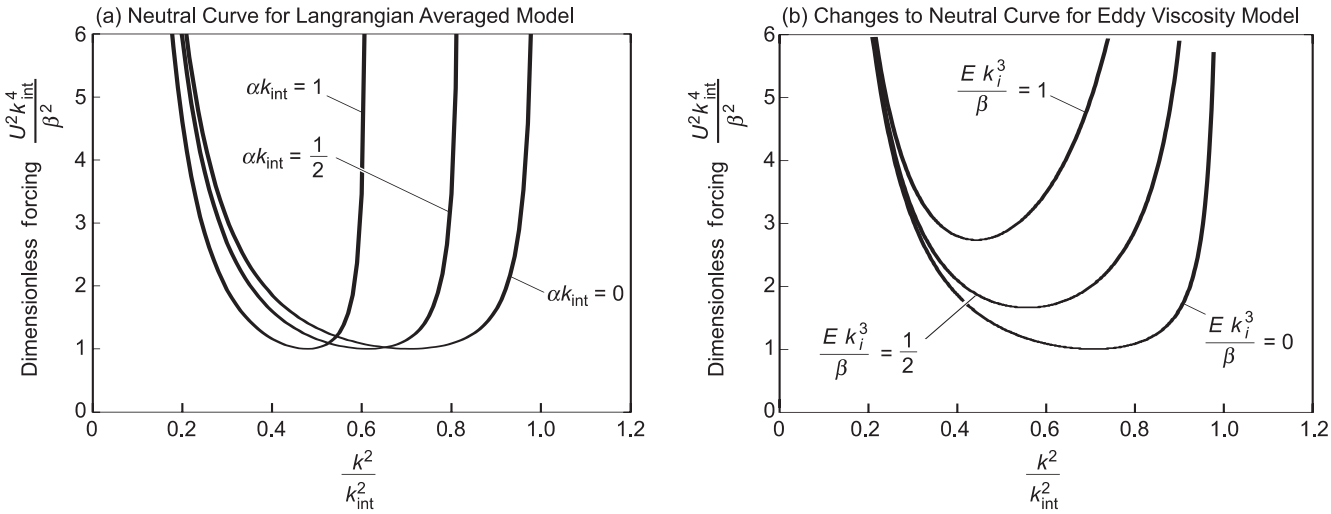


Figure 8. Baroclinic Instability: LANS-α vs Eulerian-Averaged Eddy Viscosity Models
Neutral curves show the onset of baroclinic instability (Holm and Wingate 2004). U^2 is a measure of the strength of the shear, k_{int} is the wave number of the internal Rossby deformation wave, $\beta = df/dy$, where f is the coriolis parameter, and E is the eddy viscosity. Both models lower the critical wave number for the onset of instability (the value of k at the minimum point of these neutral curves) as the modeling parameter is increased. For LANS-α, the onset occurs at the same value of the forcing irrespective of the value of α. (b) For the eddy viscosity model, the onset requires higher forcing as E is increased because of the increase in dissipation.

the eddy viscosity model for three values of Ek_{int}^3/β. As the viscosity E increases, the critical wave number again decreases, but the minimum forcing for instability gets higher rather than remaining constant. This difference arises because the LANS-α model uses dispersion to lower the critical wave number, whereas the eddy-viscosity model uses energy dissipation.The eddy-viscosity model thus requires higher forcing for the onset of baroclinic instability, and consequently some of the instability that should be present in the flow is lost as E is increased. In the LANS-α model, the gradient of potential vorticity, which drives the instability, is preserved, and therefore baroclinic instability occurs at the same forcing values as those predicted by the exact Navier-Stokes equations.

What remains unanswered is how best to choose the parameter α and how to combine both eddy-viscosity models and Lagrangian averaging in concert to achieve the most realistic results in both global ocean models and in coupled ocean, atmosphere, and ice models.

Theoretical Developments for the LANS-α Model

The Kármán-Howarth Theorem for Dynamics of the LANS-α Model. The Kármán-Howarth theorem for fluid turbulence (1938) given by Equation 7 (see box) is an exact analytical relation between the time rate of change of the second-order two-point velocity correlation function and the gradient of the third-order two-point velocity correlation function derived from the Navier-Stokes equation for homogeneous, isotropic turbulence.

Equation (7) is the lowest-order two-point statistical equation for turbulence dynamics and may be understood as a relationship between the rate of change of energy in scales of size r to the flux of energy through scales of size r.

One can write the same equation for velocity structure functions, which are the moments of the longitudinal velocity difference, $\delta_L u(\mathbf{x},t;\mathbf{r}) = \hat{\mathbf{r}} \cdot \delta\mathbf{u}(\mathbf{x},t;\mathbf{r})$, with $\delta\mathbf{u}(\mathbf{x},t;\mathbf{r}) \equiv \mathbf{u}(\mathbf{x}+\mathbf{r},t) - \mathbf{u}(\mathbf{x},t)$. One example is the second-order structure function $\langle[\delta_L u]^2\rangle$. See the articles "The Turbulence Problem" and "Direct Numerical Simulations of

The Kármán-Howarth Theorem

$$\frac{\partial}{\partial t}\langle u_i(x)u_i(x+r)\rangle - \frac{\partial}{\partial r_j}\langle u_i(x)u_j(x)u_i(x+r)\rangle = 2\nu\frac{\partial 2}{\partial r_j \partial r_j}\langle u_i(x)u_i(x+r)\rangle \quad (7)$$

where subscripts i, j denote components in a Cartesian coordinate system. The gradient of the third-order two-point velocity correlation function (the second term) arises from the nonlinear term in the Navier-Stokes equations.

Turbulence" on pages 124 and 142, respectively, for further discussion of structure functions. Kolmogorov (1941a, 1941b) used the structure function form of the Kármán-Howarth equation to show that, for homogeneous, isotropic, stationary turbulence in the limit of vanishing kinematic viscosity ($\nu \to 0$), the Navier-Stokes equations predict an exact relationship between the third-order structure function and the energy dissipation rate $\bar{\varepsilon}$ that scales linearly in the separation r namely,

$$\left\langle \left[\delta_L u\right]^3 (\mathbf{x}, t; \mathbf{r}) \right\rangle = -\frac{4}{5}\bar{\varepsilon} r \quad . \tag{8}$$

Kolmogorov's main hypotheses in deriving this relationship, which we now know as the four-fifths law were that (1) there exists an 'inertial' range of scales that are insensitive to the large flow-dependent scales and the viscous small scales, and (2) there exists a finite energy dissipation rate $\bar{\varepsilon}$ in the limit of zero viscosity. The latter is known as the dissipation anomaly for Navier-Stokes turbulence. As noted in Uriel Frisch (1995, p. 76), Kolmogorov's four-fifths law is "one of the most important results in fully developed turbulence because it is both exact and nontrivial. It thus constitutes a kind of 'boundary condition' on theories of turbulence: such theories, to be acceptable, must either satisfy the four-fifths law, or explicitly violate the assumptions made in deriving it."

Kolmogorov then assumed the self-similarity of scales in the inertial range and was able to deduce, in steps that essentially amount to dimensional analysis, that the second-order structure function must scale with $r^{2/3}$ and that, consequently, the energy spectrum (which is essentially the Fourier transform of the second-order structure function) must scale as $k^{-5/3}$.

The equivalent of the Kármán-Howarth equation was derived for the LANS-α model in Holm (2002c). Since the model relates the Helmholtz smoothed velocity **u** to the unsmoothed velocity **v**, the appropriate structure functions that emerge involve the second- and third-order two-point correlations between **u** and **v**. Upon following Kolmogorov's analysis for isotropic inertial range statistics, the corollary to the LANS-α Kármán-Howarth equation is that solutions of the LANS-α equations possess two regimes of scaling, depending on whether the separation distance r is greater, or less, than the size α. First, we find that the corresponding four-fifths law for the LANS-α model has the following scaling behavior: For $r > \alpha$, the third-order structure function $\langle[\delta u(r)]^3\rangle$ scales like r, thereby recovering Navier-Stokes behavior. In contrast, for $r < \alpha$, the third-order structure function scales like r^3. If we then assume self-similarity, we find that, for $r > \alpha$, the second-order structure function scales like $r^{2/3}$, again recovering Navier-Stokes behavior. However, for $r < \alpha$, the second-order structure function scales like r^2. Correspondingly, the power spectrum $E(k)$ for the smoothed velocity **u** has two regimes, which transition from $k^{-5/3}$ for $k\alpha < 1$ to k^{-3} for $k\alpha > 1$. Thus, the Kármán-Howarth theorem for the LANS-α model is consistent with the spectral scaling results found for it in Foias et al. (2001) by dimensional arguments.

The $k^{-5/3} \to k^{-3}$ Spectral Scaling Transition and the LANS-α Dissipation Anomaly. The LANS-α modification of Kolmogorov's four-fifths law at small separations ($r < \alpha$) results from assuming the constancy of total LANS-α energy dissipation as $\nu \to 0$. This assumption corresponds to the energy dissipation anomaly for the LANS-α model. A technical argument using embedding theorems for Besov spaces[2] implies that the LANS-α total energy dissipation is indeed constant as $\nu \to 0$ in three dimensions, provided its power spectrum $E(k)$ for kinetic energy is not steeper than k^{-4}. The k^{-3} spectrum for $k\alpha > 1$ is not too steep; therefore, the rolloff $k^{-5/3} \to k^{-3}$ in the LANS-α power spectrum is consistent with the necessary condition for possessing such an energy dissipation anomaly. Hence, the k^{-3} behavior in the power spectrum of the LANS-α model for $k\alpha > 1$ and the corresponding modification for separations $r < \alpha$ of Kolmogorov's four-fifths law derived in (Holm 2002c) are both consistent with the assumption of constant dissipation of total kinetic energy as the Reynolds number tends to infinity.

The $k^{-5/3} \to k^{-3}$ Spectral Scaling Transition and Resolution Requirements. The spectral scaling roll-off behavior for $k\alpha > 1$ has important implications for the computational performance of the LANS-α model. It substantiates the mathematical estimates of $Re^{3/2}$ for the number of degrees of freedom required for the LANS-α model to perform numerical simulations at a given Reynolds number in a periodic domain. According to this scaling, in two decades of numerical dynamic range, the LANS-α model should be able to simulate what would take three decades of numerical dynamic range for direct numerical simulation using the Navier-Stokes equations, provided the dissipation is chosen to properly balance the nonlinear transport at high wave numbers, $k\alpha >> 1$.

Implications for Smoothness of LAE-α Solutions. The r^2 behavior of the longitudinal velocity structure functions for $r < \alpha$ in the limit of zero

[2] We are grateful to G. Eyink and E. S. Titi for discussions of this argument. See Constantin and Titi (1994) and Eyink (2004) for detailed discussions.

viscosity implies the LAE-α velocity is Lipschitz continuous (Hölder index $h = 1$). That is, the velocity gradients exist almost everywhere for the LAE-α model. In contrast, the velocity for the Navier-Stokes equations in the limit of zero viscosity (the Euler equations) has Hölder index $h = 1/3$, which gives no assurance of the existence of velocity gradients for the Euler equations. On the other hand, the viscous scaling regime for the Navier-Stokes equations has Hölder index $h = 1$ and the associated r^2 scaling agrees with that found in the inviscid LAE-α model. This agreement in scaling implies that, theoretically, velocity gradients in the LAE-α model are regularized to the same degree as viscosity regularizes the gradients in the Navier-Stokes equations. Corresponding results have been verified by analytical estimates in Marsden et al. (2000). The practical implications of these theoretical results would depend, of course, on the particular numerical implementation and on other relevant parameters of a computation.

The Lagrangian-Averaged Euler-Poincaré (LAEP) Theorem. The LAEP theorem was proved by Holm (2002a, 2002b). This theorem automates the derivation of the LANS-α model and explains its relation to the generalized Lagrangian mean (GLM) theory. The GLM equations provide the exact nonlinear dynamics of Lagrangian-averaged motion, but as mentioned earlier, they are not closed. Incorporating Taylor's classic hypothesis (1921) of frozen-in Lagrangian turbulent fluctuations into the GLM equations provides the closure and yields the LAE-α model. The LANS-α model description is then obtained by introducing dissipation in the form of Navier-Stokes viscosity.

This new derivation of the LANS-α model from GLM theory and the LAEP theorem clarifies its relation to other models and shows how to extend the LANS-α model to include additional physical effects, such as rotation, buoyancy, compressibility, and magnetic fields. See Holm (2002a, 2002b) for more details.

Open Problems

Three issues have been raised in the results outlined here and in recent experience: How to understand and choose the length scale α, how to enhance our understanding of the interplay of nonlinear transport and eddy diffusion, and how to gain a more fundamental understanding of the implications of the Lagrangian-averaged fluctuation statistics of the trajectories by using data analysis.

The Length Scale α. Four heuristic interpretations for the length scale α have been proposed: (1) The size α is the length scale below which the smaller fluid circulations are swept by the larger ones and are not allowed to affect their own advection. This is the Taylor's hypothesis interpretation. (2) In the LES interpretation, the size α can be considered as a natural filter width, which defines the size of a "large" eddy in LES. (3) In its numerical interpretation, one practical rule of thumb has often been to choose α as some small integer multiple of the minimum grid spacing. In choosing α in this way, one maximizes the dynamic range left unmodified by the LANS-α model. (4) Because of its effect in slowing growth rates of instabilities at high wave numbers, Wingate (2003, Holm and Wingate 2003) suggested one could also choose the size α based on fluid and/or numerical stability requirements for numerical simulations.

All these interpretations lend heuristic insight into the physics of the particular problems we have studied using the LANS-α model. However, the length scale α is a precisely defined statistical quantity obtained from first principles. The context of Lagrangian averaging, in which the length scale α is defined, provides the basis for future developments of the LANS-α model.

Statistical Context for Future Developments. The Lagrangian statistics of the trajectory fluctuations are related to the Eulerian velocity statistics at a fixed point in space. First, the equation for the fluctuation $\mathbf{u}'$ in Eulerian velocity $\mathbf{u}(\mathbf{x},t;\,\omega) = \bar{\mathbf{u}} + \mathbf{u}'(\mathbf{x},t;\,\omega)$ for a random variable, ω, expressed in terms of the fluctuation $\xi(\mathbf{x},t;\,\omega)$ in the Lagrangian trajectory away from its mean is given by Equation (9) in the accompanying box. This relation defines the deterministic time derivative operator, D/Dt, which does not depend on the random variable ω. As a result, one finds that the exact formula for the Lagrangian dispersion tensor $\langle \xi^k \xi^l \rangle$ in terms of the Eulerian velocity statistics at a fixed point in space is given by Equation (10), where $\langle \cdot \rangle = \int (\cdot) d\mu$ now denotes average over the probability measure $d\mu$ of the random process associated with ω. The trace of this formula, given by Equation (11) in the box, is Taylor's famous dispersion law (Taylor 1921) linking the Lagrangian and Eulerian statistics of turbulence at a fixed point in space. More discussion of the role played by Taylor's contributions in the development of the LANS-α model is given in the article "Taylor's Hypothesis, Hamilton's Principle and the LANS-α Model" on page 172. The anisotropic tensor version of this formula has yet to be applied in modeling turbulence using Lagrangian statistics, and it represents an open problem in turbulence modeling.

The constant alpha case derives from Equation (9) by substituting Taylor's hypothesis that the fluctuating circulations ξ are frozen into the

Linking the Langrangian and Eulerian Statistics of Turbulence

The Eulerian velocity fluctuation $\mathbf{u}'(\mathbf{x},t;\omega)$ in terms of the Lagrangian-trajectory fluctuation $\xi(\mathbf{x},t;\omega)$ is

$$\mathbf{u}'(\mathbf{x},t;\omega) = \frac{\partial \xi}{\partial t} + \bar{\mathbf{u}} \cdot \nabla \xi - \xi \cdot \nabla \bar{\mathbf{u}} \ . \tag{9}$$

The total time derivative of the Lagrangian dispersion tensor is

$$\frac{d}{dt}\left\langle \xi^k \xi^l \right\rangle = \int \left\langle u'^k(0) u'^l(t) \right\rangle + \left\langle u'^l(0) u'^k(t) \right\rangle dt \ , \tag{10}$$

where $\langle \cdot \rangle = \int (\cdot)\, d\mu$ now denotes average over the probability measure $d\mu$ of the random process associated with ω.

The trace of Equation (10) yields Taylor's famous dispersions law linking Lagrangian and Eulerian statistics:

$$\frac{d}{dt}\left\langle |\xi|^2 \right\rangle = 2\int \left\langle \mathbf{u}'(0) \cdot \mathbf{u}'(t) \right\rangle dt \ . \tag{11}$$

Eulerian mean flow, with velocity $\bar{\mathbf{u}}$,

$$\frac{d\xi}{dt} + \bar{\mathbf{u}} \cdot \nabla \xi = 0 \ . \tag{12}$$

Hence, one finds

$$\mathbf{u}'(\mathbf{x},t;\omega) = -\xi \cdot \nabla \bar{\mathbf{u}} \ , \tag{13}$$

and, consequently,

$$\left\langle |\mathbf{u}'|^2 \right\rangle (\mathbf{x},t) = \sum_{k,l} \left\langle \xi^k \xi^l \right\rangle \left(\frac{\partial \bar{\mathbf{u}}}{\partial \mathbf{x}^k} \cdot \frac{\partial \bar{\mathbf{u}}}{\partial \mathbf{x}^l} \right) \ . \tag{14}$$

The evolution of the symmetric tensor $\langle \xi^k \xi^l \rangle$ in this formula is specified by assuming a "flow rule" for the fluctuation statistics. This is the required closure step for the Lagrangian mean theories. For example, the Taylor hypothesis—see Equation (10)—of circulations being frozen into the Eulerian mean flow implies the flow rule for the symmetric tensor,

$$\frac{d}{dt}\left\langle \xi^k \xi^l \right\rangle = 0 \ , \tag{15}$$

which preserves the initial condition that these Lagrangian statistics are homogeneous and isotropic. That is, this flow rule preserves $\langle \xi^k \xi^l \rangle = \alpha^2 \delta^{kl}$, with a constant value of α. In this case, the mean kinetic energy of the turbulent circulations simplifies to the LANS-α form,

$$\left\langle |\mathbf{u}'|^2 \right\rangle = \alpha^2 |\nabla \bar{\mathbf{u}}|^2 \ , \tag{16}$$

which relates the kinetic energy of the Eulerian velocity fluctuations to the Lagrangian statistics and the mean shear.

Other flow rules for these Lagrangian statistics possessing more sophisticated evolution equations for $\langle \xi^k \xi^l \rangle$ were catalogued in Holm (1999). However, the results of these anisotropic-tensor α equations and their comparisons with the results for the LANS-α equations in the constant alpha case have yet to be systematically explored.

Nonlinear Transport vs Diffusion and *Re* Scaling. Most of the results presented in this review depend on a trade-off between viscosity and nonlinearity in modeling the average effects of the small scales on the large ones. Consider the energy dissipated by the LANS-α equations,

$$E_\alpha = \int \left[\frac{1}{2}|\mathbf{u}|^2 + \frac{\alpha^2}{2}|\nabla \mathbf{u}|^2 \right] d^3x \ . \tag{17}$$

Following the arguments of Foias et al. (2001), the two types of energy in Equation (17) become comparable at wave number $k\alpha \approx 1$, and the scaling of the kinetic energy spectrum rolls over from $E(k) \sim k^{-5/3}$ for $k\alpha < 1$ to $E(k) \sim k^{-3}$ for $k\alpha > 1$. This change of scaling produces two different inertial regimes for the LANS-α model, depending on whether the circulations are either larger or smaller than alpha. Consequently, the modified nonlinearity in the LANS-α model shortens the inertial range relative to the inertial range for the Navier-Stokes equations. For a fixed α, the second, steeper, k^{-3} inertial range for LANS-α ends when its nonlinear transport is balanced by viscous dissipation at a wave number κ_α. The LANS-α dissipation wave number κ_α scales with the Reynolds number as $\kappa_\alpha \sim Re^{1/2}$. This scaling is to be compared with the scaling for the Kolmogorov wave number $\kappa_{Ko} \sim Re^{3/4}$, at which dissipation balances nonlinearity for the Navier-Stokes equations. Thus, the modified nonlinearity of the LANS-α model strikes a balance with viscosity at a wave number that is lower than the wave number for the Navier-Stokes equations. In turn, the new balance of the LANS-α model produces energy spectra that agree well with the spectra produced by the Navier-Stokes balance at low wave numbers ($k\alpha < 1$), but the LANS-α spectra depart from the Navier-Stokes spectra at high wave numbers ($k\alpha >> 1$), and thereby enhance the model's computability. The scaling of dissipation wave number with Reynolds is $Re^{1/2}$ for this new balance vs $Re^{3/4}$ for the Navier-Stokes

balance. This difference in scaling is the source of the improved computability for the LANS-α model.

Flow Rules for Lagrangian Statistics. The LANS-α model is, by definition, a mean field theory based on Lagrangian averaging, and Lagrangian averaging is still a young field. For example, the corresponding theory of large-deviation Lagrangian statistics for nonequilibrium processes has only recently begun to develop. New experiments and direct numerical simulations have recently begun to measure and investigate the fundamental tenets of Lagrangian trajectories in turbulence. One startling discovery in both experiments and simulations is that the Lagrangian trajectories tend to stay well localized along their mean trajectories for a long period, of the order of 30 Kolmogorov times (eddy turnover times at the dissipation scale). During this period, the Lagrangian trajectories tend to obey Taylor's hypothesis of frozen-in turbulence. Then, suddenly, large scale changes in the motion of those trajectories may occur, which apparently cause them to "forget" their previous history and start over. These experiments and simulations call for new studies of stochastic effects in Lagrangian turbulence that will take Lagrangian turbulence beyond its current status as a mean field theory. Perhaps the LANS-α model will be able to contribute as the mean field basis for these studies, and, thus, it may benefit from future achievements in this currently very active area. One potential benefit would be to include into a new generation of Lagrangian turbulence models the measured flow rules for the Lagrangian statistics that allow for the observed stochastic shifts, or punctuations, thereby occasionally and stochastically violating Taylor's deterministic hypothesis that the turbulence statistics remain frozen into the mean flow. One indication that the LANS-α model may be able to form the basis for such an interpretation is the recent discovery (Jonathan Graham, Darryl Holm, Pablo Mininni, and Annick Pouquet, private communication, November 2004) that, when magnetic fields are included, this model possesses anomalous scaling, which is the hallmark of intermittency. ■

Acknowledgments

We are enormously grateful to our friends and collaborators in this endeavor. We are especially grateful to our advisory committee, S. Y. Chen, A. J. Domaradzky, R. Donnelly, G. Eyink, U. Frisch, R. M. Kerr, S. R. Sreenivasan and E. S. Titi , as well as to the participants in our four annual turbulence workshops. We thank them for their comments, encouragement and constructive suggestions. We are also grateful to our Turbulence Working Group members, who participated with us enthusiastically in weekly meetings. Finally, we are grateful to our respective divisions at Los Alamos and to STB Division, particularly David Watkins, for leadership and encouragement in providing key opportunities to make use of the unique facilities at Los Alamos.

Further Reading

Andrews, D. G., and M. E. McIntyre. 1978. An Exact Theory of Nonlinear Waves on a Lagrangian-Mean Flow. *J. Fluid Mech.* **89**: 609.

Barenblatt, G. I., and A. J. Chorin. 1998. Scaling Laws and Vanishing Viscosity Limits in Turbulence Theory. *SIAM Rev.* **40** (2): 265.

Barenblatt, G. I., A. J. Chorin, and V. M. Prostokishin. 1997. Scaling Laws in Fully Developed Turbulent Pipe Flow. *Appl. Mech. Rev.* **50**: 413.

Bleck, R. 2002. An Oceanic General Circulation Model Framed in Hybrid Isopycnic-Cartesian Coordinates. *Ocean Modell.* **37**: 55.

Bleck, R., and L. Smith. 1990. A Wind-Driven Isopycnic Coordinate Model of the North and Equatorial Atlantic Ocean. 1. Model Development and Supporting Experiments. *J. Geophys. Res.* **95** (C3): 3273.

Chen, Q., S. Y. Chen, and G. L. Eyink. 2003. The Joint Cascade of Energy and Helicity in Three-Dimensional Turbulence. *Phys. Fluids* **15**: 361.

Chen, Q., S. Y. Chen, G. L. Eyink, and D. D. Holm. 2003. Intermittency in the Joint Cascade of Energy and Helicity. *Phys. Rev. Lett.* **90** (21): 214503.

Chen, S. Y., C. Foias, D. D. Holm, L. G. Margolin, and R. Zhang. 1999. Direct Numerical Simulations of the Navier-Stokes Alpha Model. *Physica D* **133**: 66.

Chen, S. Y., C. Foias, D. D. Holm, E. J. Olson, E. S. Titi, and S. Wynne. 1998. The Camassa-Holm Equations as a Closure Model for Turbulent Channel and Pipe Flows. *Phys. Rev. Lett.* **81**: 5338.

———. 1999a. The Camassa-Holm Equations and Turbulence in Pipes and Channels. *Physica D* **133**: 49.

———. 1999b. A Connection Between the Camassa-Holm Equations and Turbulence in Pipes and Channels. *Phys. Fluids* **11**: 2343.

Cheskidov, A. 2002. Turbulent Boundary Layer Equations. *C. R. Acad. Ser. I* **334**: 423.

Cioranescu, D., and V. Girault. 1996. Variational and Classical Solutions of a Family of Fluids of Grade Two. *C. R. Acad. Ser I* **322**: 1163.

———. 1997. Weak and Classical Solutions of a Family of Second Grade Fluids. *Int. J. Non-Linear Mech.* ***32**: 317.*

Clark, R. A., J. H. Ferziger, and W. C. Reynolds. 1979. Evaluation of Subgrid-Scale Models Using an Accurately Simulated Turbulent Flow. *J. Fluid Mech.* **91**: 1.

Constantin, P., Weinan E, and E. S. Titi. 1994. Onsager's Conjecture on the Energy Conservation for Solutions of Euler's Equation. *Commun. Math. Phys.* **165**: 207.

Danabasoglu, G., and J. C. McWilliams. 2000. An Upper-Ocean Model for Short-Term Climate Variability. *J. Climate* **13**: 3380.

Domaradzki, J. A., and D. D. Holm. 2001. Navier-Stokes-Alpha Model: LES Equations with Nonlinear Dispersion. In *Modern Simulation Strategies for Turbulent Flow*. p. 107. Edited by B. J. Geurts. Flourtown, PA: R. T. Edwards, Inc.

Domaradzki, J. A., and E. M. Saiki. 1997. A Subgrid-Scale Model Based on the Estimation of Unresolved Scales of Turbulence. *Phys. Fluids* **13**: 2148.

Dukowicz, J. K., and R. D. Smith. 1994. Implicit Free-Surface Method for the Bryan-Cox-Semtner Ocean Model. *J. Geophys. Res. – Oceans* **99** (C4): 7991.

———. 1996. Stochastic Theory of Compressible Turbulent Fluid Transport. *Phys. Fluids* **99**: 7991.

Dunn, J. E., and R. L. Fosdick. 1974. Thermodynamics, Stability, and Boundedness of Fluids of Complexity 2 and Fluids of Second Grade. *Arch. Rat. Mech. Anal.* **56**: 191.

Eyink, G. L. 2003. Local 4/5-Law and Energy Dissipation Anomaly in Turbulence. *Nonlinearity* **16**: 137.

Eyink, G. L., and K. R. Sreenivasan. 2004. Onsager and the Theory of Hydrodynamic Turbulence. To appear in *Rev. Mod. Phys.*

Fabijonas, B., and D. D. Holm. 2003. Mean Effects of Turbulence on Elliptical Instability. *Phys. Rev. Lett.* **90**: 124501.

———. 2004a. Craik-Criminale Solutions and Elliptic Instability in Nonlinear-Reactive Closure Models for Turbulence. *Phys. Fluids* **16**: 853.

———. 2004b. Multi-Frequency Craik-Criminale Solutions of the Navier-Stokes Equations. *J. Fluid Mech.* **506**: 207.

———. 2004c. Euler-Poincaré Formulation and Elliptic Instability for *n*th-Gradient Fluids. *J. Phys. A* **37**: 7609.

Foias, C., D. D. Holm, and E. S. Titi. 2001. The Navier-Stokes-Alpha Model of Fluid Turbulence. *Physica D* **152**: 505.

———. 2002. The Three Dimensional Viscous Camassa-Holm Equations, and Their Relation to the Navier-Stokes Equations and Turbulence Theory. *J. Diff. Eqs.* **14**: 1.

Frisch, U. 1995. *Turbulence— The Legacy of A. N. Kolmogorov*. Cambridge, UK: Cambridge University Press.

Gallavotti, G. 1993. Some Rigorous Results About 3D Navier-Stokes. In *Les Houches 1992 NATO-ASI. Meeting on Turbulence in Extended Systems*. p. 45. Edited by R. Benzi, C. Basdevant, and S. Ciliberto. New York: Nova Science Publishers.

Gatski, T. B., and C. G. Speziale. 1993. On Explicit Algebraic Stress Models for Complex Turbulent Flows. *J. Fluid Mech.* **254**: 59.

Gent, P. R., and J. C. McWilliams. 1990. Isopycnal Mixing in Ocean Circulation Models. *J. Phys. Oceanogr.* **20**: 150.

Gent, P. R., F. O. Bryan, G. Danabasoglu, S. C. Doney, W. R. Holland, W. G. Large, and J. C. McWilliams. 1998. The NCAR Climate System Model Global Ocean Component. *J. Climate* **11**: 1287.

Geurts, B. J., and D. D. Holm. 2002a Alpha-Modeling Strategy for LES of Turbulent Mixing. In *Turbulent Flow Computation*. p. 237. Edited by D. Drikakis and B. J. Geurts. London: Kluwer Academic Publishers.

———. 2002b. Leray Simulation of Turbulent Shear Layers. In *Advances in Turbulence IX*: *Proceedings of the Ninth European Turbulence Conference*. p. 337. Edited by J. P. Castro and P. E. Hancock. Barcelona: CIMNE.

———. 2003a. Regularization Modeling for Large-Eddy Simulation. *Phys. Fluids* **15**: L13.

———. 2003b. Commutator-Errors in Large-Eddy Simulation. Submitted to *Phys. Fluids*.

Gill, A. E. 1982. *Atmosphere-Ocean Dynamics*. San Diego, CA: Academic Press.

Gjaja, I., and D. D. Holm. 1996. Self-Consistent Wave-Mean Flow Interaction Dynamics and its Hamiltonian Formulation for a Rotating Stratified Incompressible Fluid. *Physica D* **98**: 343.

Greatbatch, R. J., and B. T. Nadiga. 2000. Four Gyre Circulation in a Barotropic Model with Double Gyre Wind Forcing. *J. Phys. Oceanogr.* **30**: 1461.

Gutmark, E., and I. Wygnanski. 1976. Planar Turbulent Jet. *J. Fluid Mech.* **73**: 465.

Holm, D. D. 1996. Hamiltonian Balance Equations. *Physica D* **98**: 379.

———. 1999. Fluctuation Effects on 3D Lagrangian Mean and Eulerian Mean Fluid Motion. *Physica D* **133**: 215.

———. 2002a. Variational Principles for Lagrangian Averaged Fluid Dynamics. *J. Phys. A* **35**: 1.

———. 2002b. Averaged Lagrangians and the Mean Dynamical Effects of Fluctuations in Continuum Mechanics. *Physica D* **170**: 253.

———. 2002c. Kármán-Howarth Theorem for the Lagrangian-Averaged Navier-Stokes-Alpha Model of Turbulence. *J. Fluid Mech.* **467**: 205.

———. 2003. Modified Speziale Model for LES Turbulence. Submitted to *J. Fluid Mech.*

Holm, D. D., and R. M. Kerr. 2002. Transient Vortex Events in the Initial Value Problem for Turbulence. *Phys. Rev. Lett.* **88** (24): 244501.

Holm, D. D., and B. Nadiga. 2003. Modeling Mesoscale Turbulence in the Barotropic Double Gyre Circulation. *J. Phys. Oceangr.* **33**: 2355.

Holm, D. D., and B. A. Wingate. 2004. Baroclinic Instability for LANS-Alpha Models. To appear in *J. Phys. Oceangr.*

Holm, D. D., J. E. Marsden, and T. S. Ratiu. 1998a. Euler-Poincaré Models of Ideal Fluids with Nonlinear Dispersion. *Phys. Rev. Lett.* **80**: 4173.

———. 1998b. Euler-Poincaré Equations and Semidirect Products with Applications to Continuum Theories. *Adv. Math.* **137**: 1.

———. 2002. The Euler-Poincaré Equations in Geophysical Fluid Dynamics. In *Large-Scale Atmosphere-Ocean Dynamics 2*: *Geometric Methods and Models*. p. 251. Edited by J. Norbury and I. Roulstone. Cambridge, UK: Cambridge University Press.

Holm, D. D., V. Putkaradze, P. D. Weidman, and B. A. Wingate. 2003. Boundary Effects on Exact Solutions of the Lagrangian-Averaged Navier-Stokes-α Equations. *J. Stat. Phys*. **113**: 841.

Horiuti, K. 2002. Roles of Nonaligned Eigenvectors of Strain-Rate and Subgrid-Scale Stress Tensors for Turbulence Generation. Submitted to *J. Fluid Mech.*

Kolmogorov, A. N. 1941a. The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Numbers. *Dok. Akad. Nauk. SSSR* **30**: 4. English translation *Proc. R. Soc. London, Ser. A* 1991, **434**: 9.

———. 1941b. Dissipation of Energy in the Locally Isotropic Turbulence. *Dok. Akad. Nauk. SSSR* **32**: 1. English translation *Proc. R. Soc. London*, *Ser. A*, 1991, **434**: 15.

Krahmann, G., and M. Visbeck. 2003. Labrador Sea Deep Convection Experiment Data Collection. Sponsored by the Office of Naval Research. [Online]: http://www.ldeo.columbia.edu/~visbeck/labsea/labsea.html

Kurien, S. 2003. The Reflection-Antisymmetric Counterpart of the Kármán-Howarth Theorem. *Physica D* **175**: 167.

Launder, B. E., and D. B. Spalding. 1974. The Numerical Computation of Turbulent Flows. *Comput. Methods Appl. Mech. Engr*. **3**: 269.

Leonard, A. 1974 Energy Cascade in Large-Eddy Simulations of Turbulent Fluid Flows. *Adv. Geophys*. **18**: 237.

Leray, J. 1934. Sur le Mouvement d'un Liquide Visqueux Emplissant L'espace. *Acta Math*. **63**: 193. Reviewed in P. Constantin, C. Foias, B. Nicolaenko and R. Temam. 1989. *Integral Manifolds and Inertial Manifolds for Dissipative Partial Differential Equations, Applied Mathematical Sciences.* Vol. 70. New York: Springer-Verlag

Lorenz, E. N. 1992. The Slow Manifold—What Is It? *J. Atmos. Sci*. **49**: 2449.

Lumley, J. L. 1970. Toward a Turbulent Constitutive Relation. *J. Fluid Mech*. **41**: 413.

Marsden, J. E., T. S. Ratiu, and S. Shkoller. 2000. The Geometry and Analysis of the Averaged Euler Equations and a New Diffeomorphism Group. *Geom. Funct. Anal.* **10**: 582.

Marsden, J. E., and S. Shkoller. 2001. Global Well-Posedness for the Lagrangian Averaged Navier-Stokes (LANS-alpha) Equations on Bounded Domains. *Philos. Trans. R. Soc. London, Ser. A* **359**: 1449.

Meneveau, C., and J. Katz. 2000. Scale-Invariance and Turbulence Models for Large-Eddy Simulation. *Annu. Rev. Fluid Mech.* **32**: 1.

Mohseni, K., B. Kosovic, J. E. Marsden, and S. Shkoller. 2001. Numerical Simulations of Forced Homogeneous Turbulence Using Lagrangian Averaged Navier-Stokes Equations.In *Proceedings of the 15th AIAA Computational Fluid Dynamics Conference*, AAIA Paper 2001-2645. Anaheim, CA: AIAA.

Mohseni, K., B. Kosovic, J. E. Marsden, S. Shkoller, D. Carati, A. Wray, and R. Rogallo. 2000. Numerical Simulations of Homogeneous Turbulence Using Lagrangian Averaged Navier-Stokes Equations. In *Center for Turbulence Research Proceedings of the 2000 Summer Program*. p. 271. NASA Ames Research Center/Stanford University.

Nadiga, B. T. 2000. Scaling Properties of an Inviscid Mean-Motion Fluid Model. *J. Stat. Phys*. **98**: 935.

Nadiga, B. T., and L. G. Margolin. 2001. Dispersive Eddy Parameterization in a Barotropic Ocean Model. *J. Phys. Oceanogr*. **31**: 2525.

Nadiga, B. T. and S. Shkoller 2001. Enhancement of the Inverse-Cascade of Energy in the Two-Dimensional Averaged Euler Equations. *Phys. Fluids* **13**: 1528.

Onsager, L. 1949. Statistical Hydrodynamics. *Nuovo Cimento (*Supplement) **6**: 279.

Pacanowski, R. C. 1995. MOM 2 Documentation, User's Guide and Reference Manual, Version 1.0x. Geophysical Fluid Dynamics Laboratory Ocean Group Technical Report 3. Geophysical Fluid Dynamics Laboratory/NOAA, Princeton University, Princeton, NJ.

Putkaradze, V., and P. Weidman. 2003. Turbulent Wake Solutions of the Prandtl-Alpha Equations. *Phys. Rev. E* **67**: 036304.

Rivlin, R. S. 1957. The Relation Between the Flow of Non-Newtonian Fluids and Turbulent Newtonian Fluids. *J. Rat. Mech. Anal.* **15**: 213.

Smagorinsky, J. 1963. General Circulation Experiments with the Primitive Equations. *Mon. Weather Rev*. **93**: 99.

*S*peziale, C. G. 1991. Analytical Methods for the Development of Reynolds-Stress Closures in Turbulence. *Annu. Rev. Fluid Mech*. **23**: 107.

———. *1*998. Turbulence Modeling for Time-Dependent RANS and VLES: A Review. *AIAA J.* **36**: 173.

Taylor, G. I. 1921. Diffusion by Continuous Movements. *Proc. London Math. Soc*. **20**: 196.

Taylor, M., S. Kurien, and G. Eyink. 2003. Recovering Isotropic Statistics in Turbulence Simulations: The Kolmogorov 4/5th-Law. *Phys. Rev. E* **68**: 026310.

Vreman, B., B. Geurts, and H. Kuerten. 1996. Large Eddy Simulation of the Temporal Mixing Layer Using the Clark Model. *Theor. Comput. Fluid Dyn.* **8**: 309.

Winckelmans, G. S., A. A. Wray, O. V. Vasilyev, and H. Jeanmart. 2001. Explicit-Filtering Large-Eddy Simulation Using the Tensor-Diffusivity Model Supplemented by a Dynamic Smagorinsky Term. *Phys. Fluids* **13**: 1385.

Wingate, B. A. 2003. Numerical Analysis of the LA Shallow Water Models. Submitted to *J. Comp. Phys.*

Zagarola, M. V. 1996. "Mean Flow Scaling of Turbulent Pipe Flow." Ph.D. thesis. Department of Mechanical and Aerospace Engineering, Princeton University.

For further information, contact Darryl D. Holm (505) 667-6398 (dholm@lanl.gov) or Beth A. Wingate (505) 665-7869 (wingate@lanl.gov).

Taylor's Hypothesis, Hamilton's Principle, and the LANS-α Model for Computing Turbulence

Darryl D. Holm

G. I. Taylor's Contributions to Lagrangian vs Eulerian Thinking about Turbulence

G. I. Taylor's Dispersion Law. An understanding of Lagrangian statistics is of great importance in the ongoing effort to develop both fundamental and practical descriptions of turbulence. For example, Prandtl's turbulent mixing length came from a Lagrangian viewpoint: It was envisioned as the turbulent analog of the mean free path of molecules in a gas. In fact, until the famous paper "Diffusion by Continuous Movements" by G. I. Taylor (1921), most turbulence theory was discussed exclusively from the Lagrangian viewpoint. However, despite the obvious importance of the Lagrangian viewpoint in turbulent combustion, reacting flows, and pollutant transport, until recently, very few measurements of Lagrangian statistics were performed at large Reynolds numbers. Instead, experimentalists performed Eulerian measurements and tried to link these measurements as best they could to the Lagrangian statistics. For example, G. I. Taylor (1921) pursued the idea originating with Prandtl and others that, "by analogy with the kinetic theory of gases," one should attempt to find ways of predicting statistical properties of the flow by taking measurements at a given point in space. One of his most influential contributions in this regard was the formula

$$\frac{d}{dt}\left\langle \left|\mathbf{X}(t)-\mathbf{X}(0)\right|^2 \right\rangle = 2\int_0^t \left\langle \mathbf{u}(0)\cdot\mathbf{u}(t)\right\rangle dt \quad . \tag{1}$$

This formula links the Lagrangian and Eulerian statistics of turbulence. In this formula, $\langle \cdot \rangle$ denotes an appropriate statistical average and the velocity $\mathbf{u}(t)$ with assumed zero mean $\langle \mathbf{u}(t)\rangle = 0$ is defined by the fundamental formula $\dot{\mathbf{X}}(t, \mathbf{X}(0)) = \mathbf{u}(\mathbf{X}(t), t)$, as a composition of functions. This is the Eulerian velocity evaluated along the Lagrangian trajectory $\mathbf{x} = \mathbf{X}(t, \mathbf{X}(0))$ whose initial position is $\mathbf{X}(t = 0, \mathbf{X}(0)) = \mathbf{X}(0)$.

Taylor's formula is actually a definition, and it is independent of the dynamics of how a real fluid moves. For example, it does not refer to the Navier-Stokes equations. However, the formula is important because it relates two different types of experimental measurements: Its left side represents the dispersion of Lagrangian traces in the types of flows that can be measured—for example, by observing how dye spreads in a turbulent flow or how a bunch of balloons disperses in the wind.[1]

In contrast, the right side of Taylor's formula can be measured by sampling the Eulerian velocity field at a single spatial location, then averaging over time, and thereby measuring its velocity correlations.

Taylor argued that the correlation function on the right (Eulerian) side of this formula specifies the statistical properties of a stationary random function, an idea which had great influence in the subsequent development of statistical treatments in turbulence theory and elsewhere. In general, the properties of the (Lagrangian) displacement would depend on the specific trajectory under consid-

eration. However, Taylor argued for assuming statistical homogeneity of the Eulerian velocities, which assumes that the stochastic process generating $\mathbf{u}(t)$ does not depend on the initial position $\mathbf{X}(0)$ of the trajectory. If, in addition, the stochastic process is statistically stationary, then so are the Eulerian velocity statistics. Thus, one reason for Taylor's formula to have been influential was that it made experimental measurements of Eulerian velocity at a single point seem relevant to turbulence. Eulerian measurements are much easier than Lagrangian measurements. Averaging the velocity at a fixed location, or comparing velocities at two fixed points in space at the same instant is much easier to perform than measuring the motion of fluid parcel trajectories carried in a chaotic flow then applying averaging techniques to them. However, Eulerian statistics are not equivalent to Lagrangian statistics, in general, and turbulence modeling must eventually deal with Lagrangian statistics.

G. I. Taylor's Microscale and Its Scaling Laws. G. I. Taylor (1921) introduced the length scale now called Taylor's microscale, which is intermediate between the integral scale L and the Kolmogorov dissipation scale η. The integral scale L contains the most energy on the average. Due to the nonlinearity of fluid dynamics, energy cascades from the integral scale down through the inertial range of smaller scales, until it reaches the Kolmogorov scale, $\eta = (\nu^3/\varepsilon)^{1/4}$, where viscous dissipation finally balances nonlinearity in the Navier-Stokes equations. Thus, Kolmogorov's dissipation scale signals the end of the inertial range, and it determines the average size of the smallest eddies, which are responsible for the energy dissipation rate ε effected by the viscosity ν. In contrast, Taylor's microscale λ is an intermediate length scale associated with energy dissipation rate, the viscosity and the Eulerian time-mean kinetic energy of the circulations $\overline{u^2}$ by Taylor's formula

$$\varepsilon = 15\nu \frac{\overline{u^2}}{\lambda^2} \quad . \tag{2}$$

G. I. Taylor (1921) argued that, dimensionally,

$$\left[\lambda^2\right] = \left[\overline{u^2}\Big/\overline{(\partial_x u)^2}\right] , \tag{3}$$

and if one assumes that viscous energy dissipation may be estimated as

$$\varepsilon \approx \overline{u^3}\Big/L = 15\nu\overline{u^2}\Big/\lambda^2 \quad , \tag{4}$$

[1] The Lagrangian statistics for the spread of such "passive tracers" was first studied quantitatively by Lewis F. Richardson (1926), in his observation of the spread of ten thousand balloons released simultaneously at the London Expo on a windy day. Each balloon contained a note asking the finder to call and tell him the location and time when the balloon came to Earth. On collecting these observations, Richardson obtained the formula,

$$\frac{d}{dt}\left\langle \left|\mathbf{X}(t)\right|^2 \right\rangle \approx \left\langle \left|\mathbf{X}(t)\right|^2 \right\rangle^{2/3} \quad ,$$

which implies the Lagrangian dispersion increases with time as $<|\mathbf{X}(t)|^2> \approx t^3$. This famous "Richardson Dispersion Law" still challenges researchers in turbulence for many reasons, not least because it shows that the dispersion properties of turbulence are "anomalous" (non-Gaussian). This is one indication of the "intermittency" of turbulence. (In contrast, ordinary diffusion due to Gaussian random motion would yield the linear time dependence $<|\mathbf{X}(t)|^2> \approx t$ for the dispersion of particles.)

$$\lambda/L \approx Re^{-1/2} \quad , \tag{5}$$

where $Re = L^{4/3}\,\varepsilon^{1/3}/\nu$ is the Reynolds number based on the integral scale. A similar estimate yields the well-known formula

$$\eta/L \approx Re^{-3/4} \tag{6}$$

for the ratio of Kolmogorov's dissipation scale to the integral scale. Thus, at a given Reynolds number Re (at the integral scale), Taylor's microscale exceeds Kolmogorov's dissipation scale by the factor

$$\lambda/\eta \approx Re^{1/4} \quad . \tag{7}$$

A physical interpretation of Taylor's microscale has recently emerged in the context of Lagrangian-averaged computational turbulence models. In particular, the LANS-α model is parameterized by the length scale α, which is the mean correlation length of a Lagrangian trajectory with its own running time average. Remarkably, the Lagrangian-averaged dynamics of the LANS-α model achieves a balance between its modified nonlinearity and its viscous dissipation, occurring at a length scale that has precisely the same Reynolds scaling as Taylor's microscale. Before explaining this result, we need to review another of Taylor's contributions linking Lagrangian statistics to the experimental interpretation of Eulerian measurements in turbulence.

G. I. Taylor's 1938 Frozen-in Turbulence Hypothesis. G. I. Taylor (1938) made the hypothesis that, because turbulence has high power at large length scales, the advection contributed by the turbulent circulations themselves must be small, compared with the advection produced by the larger integral scales, which contain most of the energy. Therefore, in such a situation, the advection of a field of turbulence past a fixed point can be taken as being mainly due to the larger, energy containing scales. This is the frozen-in turbulence hypothesis of G. I. Taylor. Although only valid when the integral scales have sufficiently high power compared with the smaller scales, this hypothesis delivered another very convenient linkage between the Eulerian and Lagrangian viewpoints of turbulence. Taylor's hypothesis holds, provided $\overline{u^2} << U^2$, where $\overline{u^2}$ is a reasonable approximation for the variations of rapidly circulating quantities that are swept along in the x-direction by the larger scales in the flow and do not influence their own evolution.

G. I. Taylor made his frozen-in turbulence hypothesis in terms of the Eulerian mean flow and, since then, others have followed suit. In experiments, this substitution allows time series measured at a single point to be interpreted as spatial variations being swept along in the Eulerian mean flow. This frozen-in turbulence advects with the Eulerian mean flow; so it remembers its initial conditions for a while. For example, advection of the three components of a vector quantity $\boldsymbol{\xi}$ by a three-dimensional Eulerian mean velocity field $\bar{\mathbf{u}}$ is expressed as

$$\frac{d}{dt}\xi\big(t,\mathbf{x}(t)\big) = \frac{\partial \xi}{\partial t} + \bar{\mathbf{u}}\cdot\nabla\xi = 0 \ , \quad \text{along} \ \frac{d\mathbf{x}}{dt} = \bar{\mathbf{u}} \ . \tag{8}$$

Thus, the advected quantity $\boldsymbol{\xi}$ remembers its initial conditions, as it is being transported by the Eulerian mean velocity of the large-scale flow. This is Taylor's hypothesis. When it holds, this hypothesis allows the very useful conversion of

data taken from single-point spatial measurements into their corresponding interpretation as temporal data, and vice versa. (Other approaches, such as two-point spatial measurements, must be used when the assumptions of Taylor's hypothesis break down.)

Using the Frozen-in Turbulence Hypothesis in a Turbulence Closure

Lagrangian averaging and the corresponding adaptation of Taylor's hypothesis of frozen-in turbulence circulations was used in Chen et al. (1998) to derive the closed system of Lagrangian-averaged Navier-Stokes-α (LANS-α) equations. This work treated the Lagrangian average of the exact flow as the large scale flow into which the turbulence circulations are frozen. Thus, Lagrangian averaging was first used to find a decomposition of the exact Navier-Stokes flow into its Lagrangian mean and rapidly circulating parts. Then Taylor's hypothesis was used as a closure approximation.

Lagrangian averaging of fluid equations is a standard technique, which is reviewed, for example, in Andrews and McIntyre (1978). However, Lagrangian averaging does not give closed equations. That is, it does not give equations expressed only in terms of Lagrangian-averaged evolutionary quantities. Something is always left over, which must be modeled when averaging nonlinear dynamics. This is because "the average of a product is not equal to the product of the averages," regardless of how one computes the averages. This difficulty is the Lagrangian-average version of the famous "closure problem" in turbulence.

The approach used in Chen et al. (1999) for deriving the closed Eulerian form of the inviscid convection nonlinearity in the LANS-α equations was based on combining two other earlier results. First, the Lagrangian-averaged variational principle of Gjaja and Holm (1996) was applied for deriving the inviscid averaged nonlinear fluid equations, which had been obtained by averaging Hamilton's principle for fluids over the rapid phase of their small turbulent circulations at a fixed Lagrangian coordinate. Second, the Euler-Poincaré theory for continuum mechanics of Holm, Marsden, and Ratiu (1998) was used for handling the Eulerian form of the resulting Lagrangian-averaged fluid variational principle. Next, Taylor's hypothesis of frozen-in turbulence circulations was invoked for closing the Eulerian system of Lagrangian-averaged fluid equations. Finally, the Navier-Stokes Eulerian viscous dissipation term was added, so that viscosity would cause diffusion of the newly defined Lagrangian-average momentum and proper dissipation of its total Lagrangian-averaged energy.

Gjaja and Holm had earlier derived (1996) a Lagrangian-average wave, mean-flow turbulent description, which allowed the turbulent circulations to propagate relative to the fluid. However, this Lagrangian-mean description was accomplished at the cost of adding complication in the form of self-consistent additional dynamical equations for the Lagrangian statistics of this type of turbulence. The use in Chen et al. (1998) of Taylor's hypothesis of frozen-in turbulence circulations simplified the description of the Lagrangian statistics, by assuming it is swept along by the Eulerian mean flow. Following the assumption that these Lagrangian statistics are homogeneous and isotropic, we and colleagues derived the new LANS-α turbulence equations with only one additional (constant) parameter, which is the length scale alpha.

According to the theory, alpha is the mean correlation length of a Lagrangian trajectory with its own running time average, at fixed Lagrangian label.

Practically speaking, the quantity alpha is the length scale in isotropic homogeneous turbulence at which the sweeping of the smaller scales by the larger ones first begins according to Taylor's hypothesis. That is, circulations at length scales smaller than alpha do not interact nonlinearly to create yet smaller ones in the process of their advection. However, these smaller circulations are fully present. In particular, their Lagrangian statistics contribute to the stress tensor, the inertial terms in the nonlinearity and the circulation theorem for the resulting LANS-α model.

Deriving the LANS-α Model

The motion equation for the LANS-α model is

$$\frac{\partial}{\partial t}\bar{\mathbf{v}} + \bar{\mathbf{u}}\cdot\nabla\bar{\mathbf{v}} + \nabla\bar{\mathbf{u}}^T\cdot\bar{\mathbf{v}} - \frac{1}{2}\nabla\left(|\bar{\mathbf{u}}|^2 + \alpha^2|\nabla\bar{\mathbf{u}}|^2\right) + \nabla\bar{p} = \nu\Delta\bar{\mathbf{v}} + \mathbf{F} \ , \tag{9}$$

with Eulerian mean velocity $\bar{\mathbf{u}}$ satisfying

$$\bar{\mathbf{v}} \equiv \bar{\mathbf{u}} - \alpha^2\Delta\bar{\mathbf{u}} \ \text{ for a constant } \alpha^2 \text{ and } \nabla\cdot\bar{\mathbf{u}} = 0 \ . \tag{10}$$

The inviscid part of this nonlinear motion equation (its left side) emerges from the Lagrangian-averaged Hamilton's principle for ideal fluids, upon using Taylor's hypothesis of frozen-in turbulence circulations. A sketch of its derivation is given below. For full details, see Holm (1999).

Hamilton's Principle for the Euler Equations. One begins with the Lagrangian $\ell\,[\mathbf{u}, D]$ in Hamilton's principle $\delta S = 0$ with $S = \int \ell[\mathbf{u}, D]dt$ for the Euler equations of incompressible fluid motion.

$$\ell[\mathbf{u}, D] = \int \frac{1}{2}D|\mathbf{u}|^2 - p(D-1)\,d^3x \ . \tag{11}$$

This Lagrangian is the kinetic energy, constrained by the pressure p to preserve the volume element $D\,d^3x$. Conservation of the volume element $D\,d^3x$, in turn, summons the continuity equation

$$\frac{d}{dt}\left(Dd^3x\right) = \left(\frac{\partial D}{\partial t} + \nabla\cdot D\mathbf{u}\right)d^3x = 0, \ \text{ along } \ \frac{d\mathbf{x}}{dt} = \mathbf{u} \ . \tag{12}$$

The constraint $D = 1$ then implies incompressibility, $\nabla\cdot\mathbf{u} = 0$, and preservation of incompressibility will determine the pressure as a Lagrange multiplier.

Varying the action yields

$$0 = \delta S = \int D\mathbf{u}\cdot\delta\mathbf{u} + \left(\frac{1}{2}|\mathbf{u}|^2 - p\right)\delta D - (D-1)\delta p\,d^3x\,dt \ . \tag{13}$$

As expected, stationarity of S under the variation of pressure δp imposes preser-

vation of volume, $D - 1 = 0$. The variations δD and $\delta\mathbf{u}$ are given in terms of arbitrary variations of the Lagrangian trajectory $\delta\mathbf{X} = \boldsymbol{\eta}(\mathbf{x}, t)$ as

$$\delta D = -\nabla \cdot D\boldsymbol{\eta} \quad \text{and} \quad \delta\mathbf{u} = \frac{\partial \boldsymbol{\eta}}{\partial t} + \mathbf{u} \cdot \nabla\boldsymbol{\eta} - \boldsymbol{\eta} \cdot \nabla\mathbf{u} \quad . \tag{14}$$

Integration by parts and use of the continuity equation yield

$$0 = \delta S = -\int D\left[\frac{\partial}{\partial t}\mathbf{u} + \mathbf{u} \cdot \nabla\mathbf{u} + \nabla\mathbf{u}^T \cdot \mathbf{u} - \nabla\left(\frac{1}{2}|\mathbf{u}|^2 - p\right)\right] \cdot \eta + (D-1)\,\delta p d^3x dt \quad . \tag{15}$$

Cancellation between the third and fourth terms finally implies Euler's equations,

$$\frac{\partial}{\partial t}\mathbf{u} + \mathbf{u} \cdot \nabla\mathbf{u} + \nabla p = 0, \quad \text{with} \quad \nabla \cdot \mathbf{u} = 0 \ , \tag{16}$$

by vanishing of the coefficient of the arbitrary vector function $\boldsymbol{\eta}$. This is the standard derivation of Euler's equations in the Euler-Poincaré theory of Holm, Marsden, and Ratiu (1998).

Hamilton's Principle for the Lagrangian-Averaged Euler α Equations. The derivation of the Lagrangian-averaged Euler-alpha (LAE-α) equations proceeds along the same lines, except one first decomposes the fluid velocity and volume element into their Eulerian mean and fluctuating parts, as

$$D = \overline{D} + D', \quad \text{and} \quad \mathbf{u} = \overline{\mathbf{u}} + \mathbf{u}' \ . \tag{17}$$

The fluctuating parts D' and $\mathbf{u}'$ of the Eulerian quantities D and $\mathbf{u}$ at a fixed point in space $\mathbf{x}$ are associated with fluctuations of the fluid parcel trajectory $\mathbf{X} = \tilde{\mathbf{X}} + \boldsymbol{\xi}(\tilde{\mathbf{X}}, t)$ around its Lagrangian mean trajectory $\tilde{\mathbf{X}}(t, \mathbf{X}_0)$. (For example, the running time average of $\mathbf{X}$ is taken at a fixed Lagrangian coordinate $\mathbf{X}_0$.) The relations between the D' and $\mathbf{u}'$ and the Lagrangian fluctuation $\boldsymbol{\xi}$, all expressed as functions of Eulerian position and time $(\mathbf{x}, t)$ are

$$D' = -\nabla \cdot (D\xi) \quad \text{and} \quad \mathbf{u}' = \frac{\partial \xi}{\partial t} + \overline{\mathbf{u}} \cdot \nabla\xi - \xi \cdot \nabla\overline{\mathbf{u}} \ . \tag{18}$$

These are linearized relations, which apply for sufficiently small fluctuations. Having used these linearized relations, we need not distinguish between Eulerian and Lagrangian averaging because the difference is only relevant at higher order in the relative amplitudes of the fluctuations. The simplest variant of the Lagrangian-averaged Euler equations is derived by substituting Taylor's hypothesis in the form

$$\frac{\partial \xi}{\partial t} + \overline{\mathbf{u}} \cdot \nabla\xi = 0, \quad \Rightarrow \quad \mathbf{u}' = -\xi \cdot \nabla\overline{\mathbf{u}} \ . \tag{19}$$

Thus, Taylor's hypothesis drastically simplifies the velocity decomposition. We now substitute this form of Taylor's hypothesis into the decomposition of fluid velocity on the Lagrangian for Euler's equations, perform the Eulerian average (in time) using the projection property $\overline{\overline{\mathbf{u}}} = \overline{\mathbf{u}}$ and then constrain the Eulerian-mean volume to be preserved $(\overline{D} - 1)$. Following these steps yields the averaged

Lagrangian

$$\bar{\ell}\left[\bar{\mathbf{u}},\bar{D}\right] = \int \frac{1}{2}\bar{D}\left(\left|\bar{\mathbf{u}}\right|^2 + \overline{\left(\xi^j\xi^k\right)}\bar{\mathbf{u}}_{,j}\cdot\bar{\mathbf{u}}_{,k}\right) - \bar{p}\left(\bar{D}-1\right)d^3x \quad . \tag{20}$$

By Taylor's hypothesis, the Lagrangian statistic $(\overline{\xi^j\xi^k})$ in this expression satisfies

$$\frac{\partial}{\partial t}\left(\overline{\xi^j\xi^k}\right) + \bar{\mathbf{u}}\cdot\nabla\left(\overline{\xi^j\xi^k}\right) = 0 \ , \tag{21}$$

upon using the projection property $\bar{\bar{\mathbf{u}}} = \bar{\mathbf{u}}$ again. Consequently, homogeneous isotropic initial conditions satisfying $(\overline{\xi^j\xi^k}) = \alpha^2\delta^{jk}$ with constant α^2 are preserved by the dynamics, and the averaged Lagrangian $\bar{\ell}[\bar{\mathbf{u}}, \bar{D}]$ in Hamilton's principle $\delta\bar{S} = 0$ with $\bar{S} = \int \bar{\ell}[\bar{\mathbf{u}}, \bar{D}]dt$ for these initial conditions simplifies to

$$\bar{\ell}\left[\bar{\mathbf{u}},\bar{D}\right] = \int \frac{1}{2}\bar{D}\left(\left|\bar{\mathbf{u}}\right|^2 + \alpha^2\left|\nabla\bar{\mathbf{u}}\right|^2\right) - \bar{p}\left(\bar{D}-1\right)d^3x \quad . \tag{22}$$

Note that the constant α^2 appears in the relative kinetic energy much the same way as Taylor argued dimensionally for his microscale. That is, α^2 encodes the relative specific kinetic energies of the Eulerian mean fluid velocity $|\bar{\mathbf{u}}|^2$ and the turbulent circulations, which satisfy $\overline{|\mathbf{u}'|^2} = \alpha^2|\nabla\bar{\mathbf{u}}|^2$ because of Taylor's hypothesis of frozen-in turbulence. However, as we shall see, α is not Taylor's microscale.

Reapplying Hamilton's variational principle with this averaged Lagrangian by following the Euler-Poincaré theory, as we did before for the Euler equations, now yields the motion equation for the Lagrangian-averaged Euler-α (LAE-α) model.

$$\frac{\partial}{\partial t}\bar{\mathbf{v}} + \bar{\mathbf{u}}\cdot\nabla\bar{\mathbf{v}} + \nabla\bar{\mathbf{u}}^T\cdot\bar{\mathbf{v}} - \frac{1}{2}\nabla\left(\left|\bar{\mathbf{u}}\right|^2 + \alpha^2\left|\nabla\bar{\mathbf{u}}\right|^2\right) + \nabla\bar{p} = 0 \ , \tag{23}$$

with

$$\bar{\mathbf{v}} \equiv \bar{\mathbf{u}} - \alpha^2\Delta\bar{\mathbf{u}} \quad \text{and} \quad \nabla\cdot\bar{\mathbf{u}} = 0 \ . \tag{24}$$

Finally, adding viscosity in Navier-Stokes form and forcing on the right side of the LAE-α model recover the LANS-α equation of motion:

$$\frac{\partial}{\partial t}\bar{\mathbf{v}} + \bar{\mathbf{u}}\cdot\nabla\bar{\mathbf{v}} + \nabla\bar{\mathbf{u}}^T\cdot\bar{\mathbf{v}} - \frac{1}{2}\nabla\left(\left|\bar{\mathbf{u}}\right|^2 + \alpha^2\left|\nabla\bar{\mathbf{u}}\right|^2\right) + \nabla\bar{p} = \nu\Delta\bar{\mathbf{v}} + \mathbf{F} \ . \tag{25}$$

Relation of LANS-α Inertial Subrange to Taylor's Microscale

The LANS-α system of equations has a variety of properties, only one of which we shall discuss here; that is, its inertial regime has two different scalings, depending on whether the circulations are either larger or smaller than alpha. In fact, its Kármán-Howarth theorem discussed in Holm (2002) implies that its kinetic energy spectrum changes from $k^{-5/3}$ for large scales, corresponding to wave numbers $k\alpha << 1$, to k^{-3} for small scales corresponding to wave numbers $k\alpha >> 1$. For a dimensional argument justifying this change of scaling in the iner-

tial regime for the LANS-α model, see Foias, Holm, and Titi (2001).

Because of this change of scaling in the LANS-α model for circulations that are larger or smaller than alpha, the inertial range is shortened for the LANS-α model. With α fixed, the wave number κ_α at the end of the second, steeper k^{-3} LANS-α inertial range is determined in Foias, Holm, and Titi (2001) to be

$$\kappa_\alpha \approx \left(\frac{1}{\alpha}\right)^{1/3} \kappa_{Ko}^{2/3} \quad . \tag{26}$$

Since the Kolmogorov dissipation wave number (κ_{Ko}) scales with integral scale Reynolds number as $\kappa_{Ko} \approx Re^{3/4}$, one finds that dissipation balances nonlinearity for the LANS-α model at $\kappa_\alpha \approx Re^{1/2}$, which is precisely the Reynolds scaling for the Taylor microscale. Thus, there is a relationship among the three progressively larger wave numbers

$$1/\alpha < \kappa_\alpha \approx Re^{1/2} < \kappa_{Ko} \approx Re^{3/4} \quad . \tag{27}$$

Shortening the inertial range for the LANS-α model to $k < \kappa_\alpha \approx Re^{1/2}$ rather than $k < \kappa_{Ko} \approx Re^{3/4}$ implies fewer active degrees of freedom in the solution for the LANS-α model, which clearly makes it much more computable than Navier-Stokes at high Reynolds numbers.

Counting Degrees of Freedom. If one expects turbulence to be "extensive" in the thermodynamic sense, then one may expect that the number of "active degrees of freedom" N_{dof} for LANS-α model turbulence should scale as

$$N_{\text{dof}}^{\alpha} \equiv \left(L\kappa_\alpha\right)^3 \approx \left(L/\alpha\right)\left(L\kappa_{Ko}\right)^2 \approx \frac{L}{\alpha} Re^{3/2} \quad , \tag{28}$$

where L is the integral scale (or domain size), κ_α is the end of the LANS-α inertial range, and $Re = L^{4/3}\varepsilon^{1/3}/\nu$ is the integral-scale Reynolds number (with total energy dissipation rate ε and viscosity ν). The corresponding number of degrees of freedom for Navier Stokes with the same parameters is

$$N_{\text{dof}}^{\text{NS}} \equiv \left(L\kappa_{Ko}\right)^3 \approx Re^{9/4} \quad , \tag{29}$$

and one sees a possible trade-off in the relative Reynolds number scaling of the two models, provided one resolves down to the Taylor microscale. (In practice, users of the LANS-α model often find acceptable results by setting its resolution scale to be just a factor of 2 smaller than α.)

Should these estimates of the number of degrees of freedom needed for numerical simulations that use the LANS-α model relative to Navier-Stokes not be overly optimistic, the implication would be a two-thirds power scaling advantage for using the LANS-α model. That is, in needing to resolve only the Taylor microscale, the LANS-α model could compute accurate results at scales larger than α by using two decades of resolution in situations that would require three decades of resolution for the Navier-Stokes equations at sufficiently high Re.

The argument for this advantage is as follows: One factor of $(N^{\text{NS}}_{\text{dof}}/N^{\alpha}_{\text{dof}})^{1/3}$

in relative increased computational speed is gained by the LANS-α model for each spatial dimension and yet another factor (at least) for the accompanying lessened Courant-Friedrichs-Levy (CFL) time step restriction. Altogether, this would be a gain in speed of

$$\left(\frac{N_{\mathrm{dof}}^{\mathrm{NS}}}{N_{\mathrm{dof}}}\right)^{4/3} = \left(\frac{\alpha}{L}\right)^{4/3} Re \quad . \tag{30}$$

Since $\alpha/L << 1$ and $Re >> 1$, the two factors in the last expressions do compete, but the Reynolds number should win out, because Re can keep increasing while the number α/L is expected to tend to a constant value, say $\alpha/L = 1/100$, at high (but experimentally attainable) Reynolds numbers, at least for simple flow geometrics. Empirical indications for this tendency were found in Chen, Foias et al. (1998, 1999a, 1999b) by comparing steady LANS-α solutions with experimental mean-velocity-profile data for turbulent flows in pipes and channels.

Thus, according to this scaling argument, a factor of 10^4 in increased speed for accurate computation of scales greater than α could occur, by using the LANS-α model at the Reynolds number for which the ratio $\kappa_{Ko}/\kappa_\alpha = 10$. An early indication of the feasibility of obtaining such factors in increased computational speed was realized in the direct numerical simulations of homogeneous turbulence reported in Chen, Holm et al. (1999), in which $\kappa_{Ko}/\kappa_\alpha \cong 4$ and the full factor of $4^4 = 256$ in computational speed was obtained using spectral methods in a periodic domain at little or no cost of accuracy in the statistics of the resolved scales. ■

For further information, contact Darryl D. Holm (505) 667-6398 (dholm@lanl.gov).

Further Reading

Andrews, D. G., and M. E. McIntyre. 1978. An Exact Theory of Nonlinear Waves on a Lagrangian-Mean Flow. *J. Fluid Mech.* **89**: 609.

Chen, S. Y., D. D. Holm, L. G. Margolin, and R. Zhang. 1999. Direct Numerical Simulations of the Navier-Stokes Alpha Model. *Physica D* **133**: 66.

Chen, S., C. Foias, D. D. Holm, E. J. Olson, E. S. Titi, and S. Wayne. 1999a. A Connection between the Camassa-Holm Equations and Turbulence in Pipes and Channels. *Phys. Fluids* **11**: 2343.

———. 1999b. The Camassa-Holm Equations and Turbulence in Pipes and Channels. *Physica D* **133**: 49.

———. 1998. The Camassa-Holm Equations as a Closure Model for Turbulent Channel and Pipe Flows. *Phys. Rev. Lett.* **81**: 5338.

Foias, C., D. D. Holm, and E. S. Titi. 2001. The Navier-Stokes-Alpha Model of Fluid Turbulence. *Physica D* **152**: 505.

Gjaja, I., and D. D. Holm. 1996. Self-Consistent Wave-Mean Flow Interaction Dynamics and Its Hamiltonian Formulation for a Rotating Stratified Incompressible Fluid. *Physica D* **98**: 343.

Holm, D. D. 1999. Fluctuation Effects on 3D Lagrangian Mean and Eulerian Mean Fluid Motion. *Physica D* **133**: 215.

———. 2002. Kármán-Howarth Theorem for the Lagrangian-Averaged Navier-Stokes-Alpha Model of Turbulence. *J. Fluid Mech.* **467**: 205.

Holm, D. D., J. E. Marsden, and T. S. Ratiu. 1998. The Euler-Poincáre Equations and Semidirect Products with Applications to Continuum Theories. *Adv. Math.* **137**: 1.

Kolmogorov, A. N. 1941. Dissipation of Energy in a Locally Isotropic Turbulence. *Dokl. Akad. Nauk SSSR* **32**: 141. (English translation in *Proc. R. Soc. London A* **434**: 15, 1991).

Richardson, L. F. 1926. Atmospheric Diffusion Shown on a Distance-Neighbour Graph. *Proc. R. Soc. London A* **110**: 709.

Taylor, G. I. 1921. Diffusion by Continuous Movements. *Proc. London Math. Soc.* **20**: 196.

———. 1938. The Spectrum of Turbulence. *Proc. R. Soc. London A* **164**: 476.

Field Theory and Statistical Hydrodynamics

The First Analytical Predictions of Anomalous Scaling

Misha Chertkov

Field theory is the most advanced subfield of theoretical physics that has been actively developing in the last 50 years. Traditionally, field theory is viewed as a formalism for solving many-body quantum mechanical problems, but the path or functional-integral representation of field theory is very useful in a much broader context. Here we discuss its use for predicting from first principles the stochastic, or turbulent, behavior of hydrodynamic flows.

The functional-integral formalism in field theory is a generalization of the famous Feynman-Kac path integral, which was introduced as a convenient alternative to the description of quantum mechanics by the Schrödinger equation. The path integral defines a quantum mechanical matrix element, or probability density for an observable, in terms of a sum over all possible trajectories, or variations, of the observable, some of which are forbidden by classical mechanics. In the more general functional-integral formalism, there is a field corresponding to each observable. In turn, to each field configuration, there is a corresponding statistical weight, and the product of the two is the integrand in the functional integral. The functional integral, which constitutes a summation or integration over many realizations of that field or observable, provides the probability distribution function for the observable.

In the 1950s, many researchers understood that any problem involving random variables, or random fields, can be interpreted in terms of a sum or integral over many field trajectories or field configurations. The simplest example is the problem of diffusion. There, the probability distribution function for the distance traveled by a single molecule as it collides at random with other molecules in a medium is calculated as a path integral over many Brownian motion trajectories. The integral reformulation is often advantageous because it allows one to utilize very powerful theoretical tools to evaluate or approximate the integrals. Perturbative analysis (often formulated in terms of diagrammatic techniques), saddle-point, or instanton, techniques, and various transformations (change of integration variables) are among the most useful tricks that allow analytical or semianalytical (numerical evaluation follows a theoretical step) evaluations.

Any problem in turbulence, or for that matter any statistical problem, formulated in terms of stochastic ordinary or partial differential equations can be reformulated in terms of a functional, or path, integral. Many researchers have contributed to the development of this field-theoretical approach to the problem of turbulent flow. This approach is now called statistical hydrodynamics. In the 1960s the most notable contributions to statistical hydrodynamics came from Robert Kraichnan (1967), who discovered the idea of the inverse cascade for two-dimensional (2-D) turbulence, and from Vladimir Zakharov (1967), who put the

theory of wave turbulence on a firm mathematical ground by finding turbulence spectra as exact solutions and by introducing the notion of inverse and dual cascades in wave turbulence.

Perturbative (or diagrammatic) analysis, which was at the core of the Kraichnan-Zakharov analysis, defined the spirit of the most important theoretical results in statistical hydrodynamics for some 30 years following publication of Kraichnan and Zakharov's seminal papers cited above. The work described in those papers was cited in the award write-up for the 2003 Dirac Medal that went to Kraichnan and Zakharov.[1]

Between 1994 and 1995, however, three independent groups (refer to Chertkov et al. 1995, Chertkov and Falkovich 1996, Gawedzki and Kupiainen 1995, Shraiman and Siggia 1995, Pumir et al. 1997) came to the conclusion that the perturbative approach, which apparently led to self-similar scaling laws for the correlation and other structure functions of Navier-Stokes turbulence, did not work for passive scalar turbulence. By applying nonperturbative field-theoretic techniques, these groups were able to prove the existence of anomalous scaling in passive scalar turbulence. Below, we outline these new developments and discuss the possible implications for anomalous scaling in both theoretical and applied contexts of turbulent flow.

Intermittency and the Passive Scalar Model

Passive scalar turbulence describes the advection and diffusion of a scalar quantity (such as temperature or pollutant concentration) in a turbulent flow. The scalar quantity is described by a scalar field $\theta(t,\mathbf{r})$, and the dynamics of the scalar field evolve in space $\mathbf{r}$ and time t according to the following linear equation:

$$\partial_t\theta + (\mathbf{u}\nabla)\theta = \kappa\Delta\theta + \phi \ , \tag{1}$$

where κ, $\mathbf{u}(t,\mathbf{r})$ and $\phi(t,\mathbf{r})$ stand for the diffusion (either thermal or material) coefficient, the incompressible velocity field, and the source field controlling injection of the scalar θ, respectively. The advection of θ is a passive process under the assumption that all three fields—velocity $\mathbf{u}$, injection ϕ, and scalar θ—are statistically independent of each other. That assumption, which is realistic in many practical cases, means that effects of the scalar field fluctuations on the flow (for example, buoyancy) are neglected.

A few years after Kolmogorov (1941) proposed the inertial cascade four-fifths law, relating third moment of velocity increment to the energy flux and energy dissipation in Navier-Stokes turbulence, Obukhov (1949) and Corrsin (1951) independently suggested that a similar consideration applies to the passive scalar problem. Indeed, if diffusion and injection are removed from Equation (1), then the integral of θ^2 over all space, $\int d\mathbf{r}\theta^2$, is conserved (or does not change with time). One can therefore consider θ^2 in the passive scalar problem as analogous to kinetic energy density, or $\mathbf{u}^2$, in Navier-Stokes turbulence. In any turbulent flow, the velocity fluctuations grow with scale size in the inertial range of scales, which lies between the dissipation scale η and the large forcing scaling L. Analogously, if the diffusion coefficient κ is small while the source field ϕ injects the "scalar energy" at a relatively large scale, L_ϕ, then advection dominates diffusion in the so-called convection range, which extends from L_ϕ down to the diffu-

sive scale, r_d. The ratio of the two scales L_ϕ/r_d is a large dimensionless number that plays a role in passive scalar turbulence analogous to the role of the pumping-to-viscous scale ratio in the Navier-Stokes turbulence. That is, when the dimensionless ratio L_ϕ/r_d (closely related to the Peclet/Schmidt numbers) becomes large, passive scalar turbulence develops.

In the Obukhov-Corrsin picture of the passive scalar problem, once a large blob of the scalar field (that is, large on the scale of L_ϕ) is injected into a turbulent flow, turbulent advection causes a fine spatial structure of scalar inhomogeneities to develop within the initially homogeneous cloud. The finest scale of the spatial inhomogeneities is r_d because inhomogeneities at even smaller scales are smeared out by diffusion. In the language of the θ^2-energy "budget," the scalar energy density θ^2, which is permanently supplied at the large scale L_ϕ, cascades toward smaller scales within the convective range and is dissipated at the small scales, approximately r_d. Thus, the analog of Kolmogorov's four-fifths law for the scalar energy flux in passive scalar turbulence reads

$$\langle \theta_1 \mathbf{u}_2 \theta_2 \rangle = -\varepsilon_\phi r_{12} \quad , \tag{2}$$

where $\langle \ldots \rangle$ describes averaging, with respect to statistics, of both velocity and injection fields and ε_ϕ is the averaged scalar-energy dissipation rate, $\varepsilon_\phi = \kappa \langle (\nabla\theta)^2 \rangle$. In this Obukhov-Corrsin picture, the flux of θ^2 remains constant from scale to scale within the convective range, and the scalar-energy dissipation rate is equal to the scalar-energy input rate at the injection scale, estimated as $\varepsilon_\phi \sim \theta^2_{L_\phi} u_{L_\phi}/L_\phi$, where u_{L_ϕ} and θ_{L_ϕ} are typical values of velocity and scalar fluctuations at the injection scale.

Equation (2), which is the passive scalar analog of the four-fifths law controlling the scalar energy budget, is exact. The exact statement, however, is limited to the very special correlation function, and no generalization is known of Equation (2) for other simultaneous correlation functions of the scalar field. This caveat was "fixed" by Obukhov and Corrsin, who conjectured self-similarity of scalar fluctuations. The conjecture is akin to Kolmogorov's self-similarity assumption for velocity fluctuations.

The self-similarity for the scalar-field statistics looks simple and thus appealing. However, accurate experimental measurements between the 1960s and the 1980s (Sreenivasan 1991, Sreenivasan and Antonia 1997), supported neither the Kolomogorov nor the Obukhov-Corrsin predictions for self-similar scaling laws, thus offering an early hint that anomalous scaling is common in turbulence. For the passive scalar increments, the anomalous scaling scenario means that the moments of scalar increments have the following form:

$$\left\langle \left[\theta(r) - \theta(r+l)\right]^{2n} \right\rangle \sim \frac{\varepsilon_\phi^n}{\varepsilon^{n/3}} l^{4n/3} \left(\frac{L_\phi}{l}\right)^{\Delta_{2n}} \quad , \tag{3}$$

where $\Delta_{2n} > 0$ is the anomalous exponent. In this formal description, the self-similar scenario would correspond to $\Delta_{2n} = 0$. The anomalous scaling, and thus lack of self-similarity, appeared to be much more pronounced in the experimental data for the scalar field than for the velocity field. Because at that time there was no theoretical understanding of the origin of anomalous scaling, the observations were essentially rejected as spurious.

Resolution of the standoff on anomalous scaling emerged in the mid-1990s. First, Kraichnan proposed (1994) an ad hoc scheme for producing a closed set of equations for what is today called the Kraichnan model. This microscopic model,

initially introduced in 1967 (refer to Kraichnan 1967), deals with passive scalar turbulence for a velocity field in Equation (1) that has self-similar statistics. The velocity field in the model was chosen to be incompressible, Gaussian, and short correlated (δ-correlated) in time. Spatial correlations in the model are characterized by the pair correlation function of the velocity difference between two points measured at two distinct times:

$$\left\langle \left(v^{\alpha}(t_1;r_1) - v^{\alpha}(t_1;r_2) \right) \left(v^{\beta}(t_2;r_1) - v^{\beta}(t_2;r_2) \right) \right\rangle = \delta(t_1 - t_2) K^{\alpha\beta}(r_1 - r_2) \tag{4}$$

where $\alpha, \beta = 1, \cdots, d$. The eddy diffusivity tensor $K^{\alpha\beta}(r)$ is growing algebraically with the spatial separation $K \propto r^{2-\gamma}$ so that the exponent characterizing the degree of non-smoothness of the synthetic velocity field γ and the spatial dimensionality d are two independently controlled parameters. In his 1994 paper, Kraichnan proposed an approximate closure scheme resulting in a closed set of equations for scalar structure functions of order 4, $S_4(l) = \langle [\theta(\mathbf{r} + \mathbf{1}) - \theta(\mathbf{r})]^4 \rangle$, and higher. The main message here was that, although the velocity field exhibited self-similarity, the scalar fluctuations are extremely intermittent and thus characterized by an anomalous expression generalizing Equation (3)

$$\left\langle \left[\left(\theta(r+l) - \theta(r) \right]^{2n} \right\rangle \sim l^{n\xi_2} \left(\frac{L_\phi}{l} \right)^{\Delta_{2n}} \propto l^{\xi_{2n}} \tag{5}$$

with $\xi_{2n} \neq n\xi_2$ and $\Delta_{2n} \neq 0$. Then, independently, and almost simultaneously, three groups (refer to Chertkov et al. 1995, Chertkov and Falkovich 1996, Gawedzki and Kupiainen 1995, Shraiman and Sigia 1995, Pumir et al. 1997) developed a rather different approach that required no ad hoc closure assumptions.

The new approach focused on the analysis of the simultaneous correlation function of the scalar field taken at four different points, $F_{1234} \equiv \langle \theta(t,\mathbf{r}_1)\, \theta(t,\mathbf{r}_2)\, \theta(t,\mathbf{r}_3)\, \theta(t,\mathbf{r}_4) \rangle$. That four-point correlation function is governed by a second-order linear, and therefore closed (!!!), partial differential inhomogeneous equation,

$$\hat{L} F_{1234} = \chi \ , \tag{6}$$

where

$$\hat{L} \equiv \sum_{i,j} K^{\alpha\beta}(r_i - r_j) \nabla_i^{\alpha} \nabla_j^{\beta}$$

is the differential operator of the second order, called eddy diffusivity operator, and χ is a known function, so that no ad hoc closure was required. The solution of any linear differential equation can be presented for a subinternal range of scales as a sum of homogeneous and certain inhomogeneous solutions of the equation. (For the four-point correlation function, the subinternal range would be the convective range of scales in which the separations between the four points are larger than the diffusive scale but smaller than the scalar injection scale.) Progress came from the recognition that the anomalous scaling contributions to the four- through n-point correlation functions, and respectively to the fourth- through nth-order moments of the scalar increments (that is, structure functions), originate primarily from homogeneous solutions of the partial differential equation, that is, from the zero modes Z of the eddy diffusivity operator $LZ = 0$. Thus,

the first important outcome of the analysis was that the value of the anomalous exponent for the passive scalar structure functions was insensitive to the strength of the forcing field. It was also shown that the anomalous contribution originates from matching the homogeneous and inhomogeneous solutions at the injection scale rather than the diffusive scale. Zero modes of the eddy diffusivity operator were analyzed and anomalous corrections Δ_{2n} were calculated in some important limits of the Kraichnan model corresponding to (a) a high spatial dimension, $d \to \infty$, so that calculations were done in an expansion with respect to $1/d$ (Chertkov et al 1995, Chertkov and Falkovich 1996), (b) an extremely irregular (diffusive) velocity, $2 - \gamma << 1$ (Gawedzki and Kupiainen 1996), (c) an almost spatially smooth velocity, $\gamma << 1$ (Shraiman and Siggia 1995, Pumir et al. 1997), and later for (d) a large deviation, or instanton, regime for which it was shown that the structure function exponent ξ_{2n} saturates to a constant (Chertkov 1997, Balkovsky and Lebedev 1998). For the first time ever, analytical calculations of a turbulence problem predicted the existence and the degree of anomalous scaling.

Passive transport in general and anomalous scaling in particular have also been given a transparent Lagrangian interpretation: It was shown that the n-point Eulerian (simultaneous) correlation function can be reinterpreted in terms of Lagrangian trajectories of n particles/markers evolving in the same velocity field. Thus, the Eulerian pair-correlation function of the scalar field $\langle \theta(\mathbf{r}+\mathbf{1})\theta(\mathbf{r})\rangle$ is equal to the value of the θ^2 energy flux ε_ϕ multiplied by the time $\langle T_{l\to L_\phi}\rangle$, which is defined as the average (over velocity field statistics) of the time for two particles released a distance r_{12} apart to become separated by a distance L_ϕ. In this Lagrangian interpretation, the anomalous scaling is related to correlations between Lagrangian trajectories of different particles—for example, $\langle T_{r_{12}\to L} T_{r_{34}\to L}\rangle \neq \langle T_{r_{12}\to L}\rangle\langle T_{r_{34}\to L}\rangle$. (That is, two pairs of particles, 1-2 and 3-4 respectively, released in the same flow diverge so that both r_{12} and r_{34} reach the integral scale L in finite times, $T_{r_{12}\to L}$ and $T_{r_{34}\to L}$, respectively. However, if the gedanken experiment is repeated many times, one finds that the two times are actually correlated; that is, they are statistically dependent. The Lagrangian interpretation of passive scalar transport has also allowed efficient numerical analysis of the problem (Frisch et al. 1998), leading to accurate validation of the theoretical results but, more important, to a wide exploration of anomalous scaling in the intermediate parametric region—away from the asymptotic limits considered in Chertkov et al. (1995), Chertkov and Falkovich (1996), Gawedski and Kupiainen (1995), Shraiman and Siggia (1995), Pumir et al. (1997), Chertkov (1997) and Balkovsky and Lebedev (1998)—where quantitative theoretical analysis had been hopeless.

In an independent development, Burgers turbulence (or simply "Burgulence") was found to have anomalous scaling of an extreme kind: The left (negative) values' tail of the probability distribution function for the velocity increment is of extremely extended, algebraic form (Chekhlov and Yakhot 1995, Polyakov 1995, Khanin et al. 1997, Frisch and Bec 2001).

These nonperturbative results on anomalous scaling in relatively simple problems are recognized as the most important breakthrough in the theory of turbulence for the following reasons: (1) They prove that anomalous scaling as an extreme form of intermittency does exist. They also demonstrate that anomalous scaling is a generic phenomenon. Now, rather than proving the existence of anomalous scaling, the major task is to explain why the anomalous scaling exponent is so small (although still distinguishable from zero) in many more complex situations such as isotropic homogeneous Navier-Stokes turbulence. (2) The new nonperturbative approach has benefited from a Lagrangian description. Thus, in

the passive scalar case, differential equations for scalar correlation functions can be reinterpreted in terms of a path integral over many Lagrangian trajectories (each set of trajectories corresponding to a single realization of velocity field). (3) The development of scalar turbulence theory (Shraiman and Siggia 2000, Falkovich et al 2001) has also generated new results in related areas of research such as kinematic dynamo theory (Vergassola 1996, Chertkov et al. 1999), enhancement of chemical reactions by turbulence (Chertkov 1999, Chertkov and Lebedev 2003), polymer stretching by turbulence (Balkovsky et al. 2000 and 2001; Chertkov 2000), elastic turbulence (Fouxon and Lebedev 2003), and more.

The progress achieved in scalar turbulence has also generated a resurgence of interest in more complex problems in statistical hydrodynamics. Motivated by the Lagrangian representation of passive scalar transport, we and colleagues have found a finite number of Lagrangian particles (four, or a tetrad, is the minimum number—see Chertkov et al. (1999) can be considered a sensible closure framework for a Lagrangian phenomenological model of Navier-Stokes turbulence. Finally, the two solvable models have opened possibilities for benchmarking various nonperturbative methods of statistical hydrodynamics such as instanton calculus (Chertkov 1997, Balkovsky and Lebedev 1998, Falkovich et al. 1996, Balkovsky et al. 1997). Our optimistic expectation is that these powerful theoretical methods may soon deliver new results for more complex and challenging problems in statistical hydrodynamics, including homogeneous isotropic Navier-Stokes turbulence, shear-driven turbulence, and perhaps even Rayleigh-Taylor turbulent mixing and magnetohydrodynamic turbulence. ■

Further Reading

Balkovsky, E., and V. Lebedev. 1998. Instanton for the Kraichnan Passive Scalar Problem. *Phys. Rev. E* **58** (5): 5776.

Balkovsky, E., A. Fouxon, and V. Lebedev. 2000. Turbulent Dynamics of Polymer Solutions. *Phys. Rev. Lett.* **84** (20): 4765.

———. 2001. Turbulence of Polymer Solutions. *Phys. Rev. E* **64**: 056301.

Balkovsky, E., G. Falkovich, I. Kolokolov, and V. Lebedev. 1997. Intermittency of Burgers' Turbulence. *Phys. Rev. Lett.* **78** (8): 1452.

Bernard, D., K. Gawedzki, and A. Kupiainen. 1996. Anomalous Scaling in the *N*-Point Functions of a Passive Scalar. *Phys. Rev. E* **54** (3): 2564.

Chekhlov, A., and V. Yakhot. 1995. Kolmogorov Turbulence in a Random-Force-Driven Burgers Equation. *Phys. Rev. E* **51** (4): R2739.

Chertkov, M. 1997. Instanton for Random Advection. *Phys. Rev. E* **55** (3): 2722.

———. 1998. On How a Joint Interaction of Two Innocent Partners (Smooth Advection and Linear Damping) Produces a Strong Intermittency. *Phys. Fluids* **10** (11): 3017.

———. 1999. Passive Advection in Nonlinear Medium. *Phys. Fluids* **11** (8): 2257.

———. 2000. Polymer Stretching by Turbulence. *Phys. Rev. Lett.* **84** (20): 4761.

Chertkov, M., and G. Falkovich. 1996. Anomalous Scaling Exponents of a White-Advected Passive Scalar. *Phys. Rev. Lett.* **76** (15): 2706.

Chertkov, M., and V. Lebedev. 2003a. Boundary Effects on Chaotic Advection-Diffusion Chemical Reactions. *Phys. Rev. Lett.* **90** (13): 134501.

———. 2003b. Decay of Scalar Turbulence Revisited. *Phys. Rev. Lett.* **90** (3): 034501.

Chertkov, M., A. Pumir, and B. I. Shraiman. 1999. Lagrangian Tetrad Dynamics and the Phenomenology of Turbulence. *Phys. Fluids* **11** (8): 2394.

Chertkov, M., G. Falkovich, I. Kolokolov, and V. Lebedev. 1995. Normal and Anomalous Scaling of the Fourth-Order Correlation Function of a Randomly Advected Passive Scalar. *Phys. Rev. E* **52** (5): 4924.

Chertkov, M., G. Falkovich, I. Kolokolov, and M. Vergassola. 1999. Small-Scale Turbulent Dynamo. *Phys. Rev. Lett.* **83** (20): 4065.

Corrsin, S. 1951. On the Spectrum of Isotropic Temperature Fluctuations in an Isotropic Turbulence. *J. Appl. Phys.* **22** (4): 469.

E, Weinan, K. Khanin, A. Mazel, and Y. Sinai. 1997. Probability Distribution Functions for the Random Forced Burgers Equation. *Phys. Rev. Lett.* **78** (10): 1904.

Falkovich, G., K. Gawedzki, and M. Vergassola. 2001. Particles and Fields in Fluid Turbulence. *Rev. Mod. Phys.* **73**: 913.

Falkovich, G., I. Kolokolov, V. Lebedev, and A. Migdal. 1996. Instantons and Intermittency. *Phys. Rev. E* **54** (5): 4896.

Fouxon, A., and V. Lebedev. 2003. Spectra of Turbulence in Dilute Polymer Solutions. *Phys. Fluids* **15** (7): 2060.

Frisch, U., and J. Bec. 2001. "Burgulence". In *Les Houches Session LXXIV. New Trends in Turbulence. Turbulence: Nouveux Aspects.* Edited by M. Lesieur, A. Yaglom, and F. David. (Grenoble, France, 2000), p. 341.

Frisch, U., A. Mazzino, and M. Vergassola. 1998. Intermittency in Passive Scalar Advection. *Phys. Rev. Lett.* **80** (25): 5532.

Gawedzki, K., and A. Kupiainen. 1995. Anomalous Scaling of the Passive Scalar. *Phys. Rev. Lett.* **75** (21): 3834.

Kolmogorov, A. N. 1941. The Local Structure of Turbulence in Incompressible Viscous Fluid for Very Large Reynolds Numbers. *C. R. Acad. Sci. USSR* **30**: 301.

Kraichnan, R. H. 1994. Anomalous Scaling of a Randomly Advected Passive Scalar. *Phys. Rev. Lett.* **72** (7): 1016.

Kraichnan, R. H. 1967. Inertial Ranges in 2-Dimensional Turbulence. *Phys. Fluids* **10** (7): 1417.

———. 1975. Statistical Dynamics of 2-Dimensional Flow. *J. Fluid Mech.* **67**: 155.

———. 1971. Inertial-Range Transfer in 2-Dimensional and 3-Dimensional Turbulence. *J. Fluid Mech.* **47**: 525.

Obukhov, A. M. 1949. Structure of the Temperature Field in Turbulence. *Izv. Acad. Nauk. USSR, Ser. Geogr. Geophys.* **13**: 55.

Polyakov, A. M. 1995. Turbulence without Pressure. *Phys. Rev. E* **52** (6): 6183.

Pumir, A., B. I. Shraiman, and E. D. Siggia. 1997. Perturbation Theory for the δ-Correlated Model of Passive Scalar Advection Near the Batchelor Limit. *Phys. Rev. E* **55** (2): R1263.

Shraiman, B. I., and E. D. Siggia. 1995. Anomalous Scaling of a Passive Scalar in Turbulent Flow. *C. R. Acad. Ser. II* **321** (7): 279.

Shraiman, B. I., and E. D. Siggia. 2000. Scalar Turbulence. *Nature* **405**: 639.

Sreenivasan, K. R. 1991. On Local Isotropy of Passive Scalars in Turbulent Shear Flows. *Proc. R. Soc. London, Ser. A* **434**: 165.

Sreenivasan, K. R., and R. A. Antonia. 1997. The Phenomenology of Small-Scale Turbulence. *Annu. Rev. Fluid Mech.* **29**: 435.

Vergassola, M. 1996. Anomalous Scaling for Passively Advected Magnetic Fields. *Phys. Rev. E* **53** (4): R3021.

Zakharov, V. E. 1967. Weak-Turbulence Spectrum in a Plasma without a Magnetic Field. *Sov. Phys. JETP* **24** (2): 455.

Zakharov, V. E., V. S. Lvov, and G. Falkovich. 1992. *Kolmogorov Spectra of Turbulence*. Berlin; New York: Springer-Verlag.

For further information, contact Misha Chertkov (505) 665-8119 (chertkov@lanl.gov).

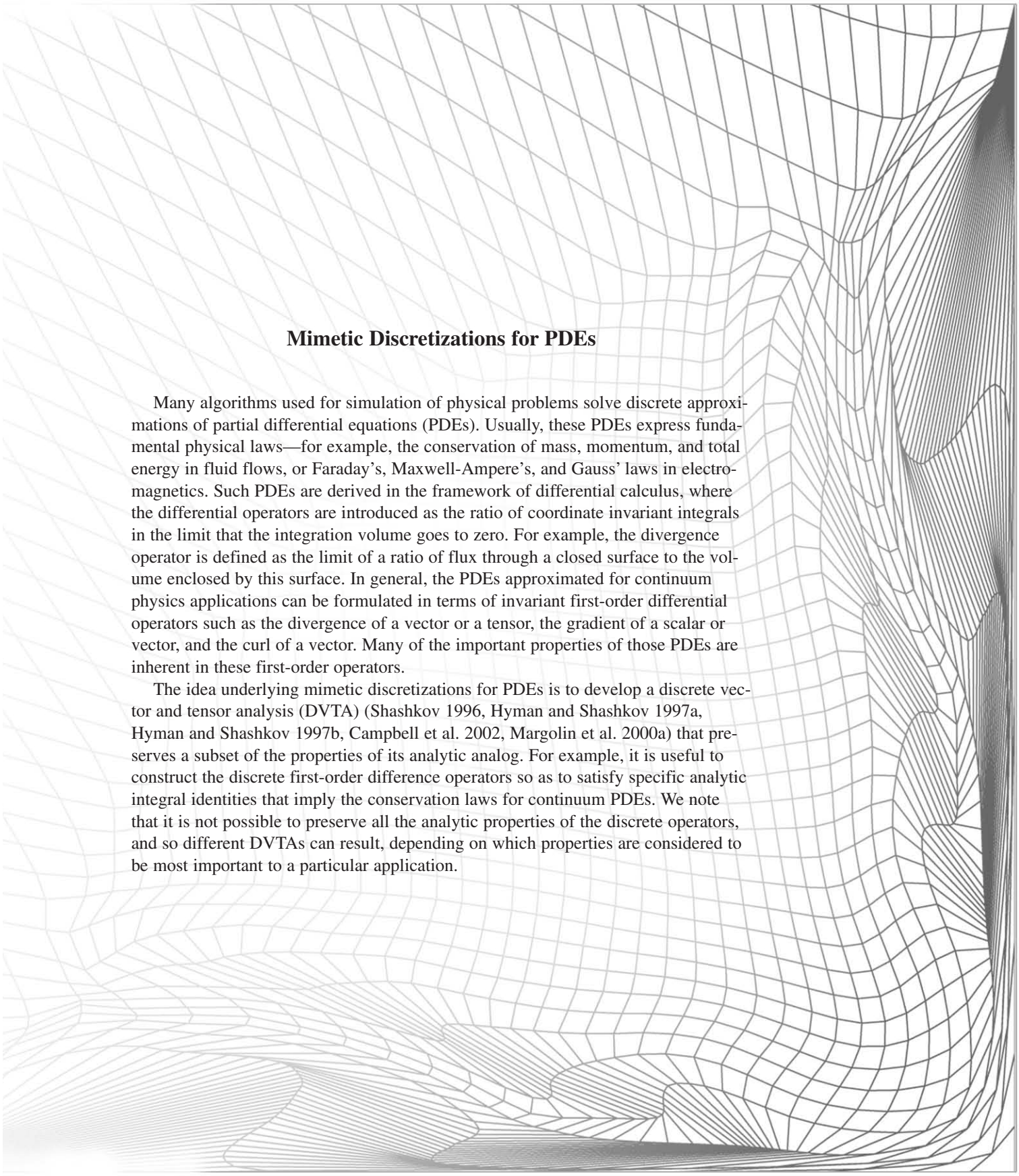
Physically Motivated Discretization Methods

A Strategy for Increased Predictiveness

Dana Knoll, Jim Morel, Len Margolin, and Misha Shashkov

Los Alamos is one of the birthplaces of computational science. The need of the weapons program to approximate the solutions of strongly nonlinear, coupled partial differential equations in complex domains has been a continuous driver in the dual development of supercomputing platforms and of more accurate and efficient numerical algorithms. More recently, the cessation of nuclear testing has placed a new requirement on algorithms, that of increased predictiveness.

Despite the importance and magnitude of the effort that has been put into computational science, in many ways the construction of new algorithms remains more of an art than a science. While the accuracy and efficiency of an algorithm can be studied and enhanced with the mathematical tools of numerical analysis, increased predictiveness is more typically the result of incorporating physical principles into the algorithm. In this article, we describe three examples of methodologies for improving predictiveness of numerical simulations: mimetic differencing, asymptotic-preserving discretization, and implicitly balanced solution techniques. The first two methodologies are focused on spatial discretization, and the third, on temporal discretization. Each is attempting to embed some basic underlying physical concept into the numerical method, thereby improving the fidelity and predictive capability of computer simulation. At some level, these methodologies are currently being incorporated in existing or next-generation simulation software within the Los Alamos weapons program.



Mimetic Discretizations for PDEs

Many algorithms used for simulation of physical problems solve discrete approximations of partial differential equations (PDEs). Usually, these PDEs express fundamental physical laws—for example, the conservation of mass, momentum, and total energy in fluid flows, or Faraday's, Maxwell-Ampere's, and Gauss' laws in electromagnetics. Such PDEs are derived in the framework of differential calculus, where the differential operators are introduced as the ratio of coordinate invariant integrals in the limit that the integration volume goes to zero. For example, the divergence operator is defined as the limit of a ratio of flux through a closed surface to the volume enclosed by this surface. In general, the PDEs approximated for continuum physics applications can be formulated in terms of invariant first-order differential operators such as the divergence of a vector or a tensor, the gradient of a scalar or vector, and the curl of a vector. Many of the important properties of those PDEs are inherent in these first-order operators.

The idea underlying mimetic discretizations for PDEs is to develop a discrete vector and tensor analysis (DVTA) (Shashkov 1996, Hyman and Shashkov 1997a, Hyman and Shashkov 1997b, Campbell et al. 2002, Margolin et al. 2000a) that preserves a subset of the properties of its analytic analog. For example, it is useful to construct the discrete first-order difference operators so as to satisfy specific analytic integral identities that imply the conservation laws for continuum PDEs. We note that it is not possible to preserve all the analytic properties of the discrete operators, and so different DVTAs can result, depending on which properties are considered to be most important to a particular application.

The construction of a mimetic discretization for a particular PDE starts with the choice of a discrete representation of the scalar and vector fields—what is usually termed the data structure. (Here, we are considering discretizations that employ a computational mesh, which is the most common but by no means the only choice.) For example, in electromagnetics it is natural to choose the normal projections of magnetic flux density with respect to the faces of the computational cells and the normal projections of electric field intensity to edges of the computational cells as primary variables, because these components of the magnetic and electric fields are continuous at an interface between different materials (Hyman and Shashkov 1999a). On the other hand, in Lagrangian gas dynamics, it is natural to locate the Cartesian components of velocity at the nodes of the mesh because, in a Lagrangian framework, the nodes of the mesh move with the fluid (Caramana et al. 1998b).

The next step is to identify the connection between the most significant properties of the model PDEs and the first-order differential operators in terms of which they are written. For example, the conservation of total energy in Lagrangian gas dynamics formally follows from the property that the analytic gradient operator is the negative adjoint of the analytic divergence operator (Shashkov 1996):

$$\int_V p\nabla\cdot\mathbf{W}dV + \int_V \mathbf{W}\cdot\nabla p dV = \oint_{\partial V} p\mathbf{W}\cdot\mathbf{n}dS \quad , \tag{1}$$

where p is the (scalar) pressure and $\mathbf{W}$ is the (vector) velocity field. Similarly,the conservation of momentum in the equations of gas dynamics follows from the following property of the gradient:

$$\int_V \nabla p dV = \oint_{\partial V} p\mathbf{n}dS \quad . \tag{2}$$

A third example arises in solid dynamics, where the velocity derivatives are used to estimate the strain-rate tensor. Here it is important to define the discrete divergence operator so that the divergence of velocity is consistent with the change of volume of a material parcel (Margolin et al. 2000a):

$$\nabla\cdot\mathbf{W} = \lim_{\delta V\to 0}\frac{\frac{d}{dt}(\delta V)}{\delta V} \quad . \tag{3}$$

Sometimes it is not possible to formulate discrete operators that satisfy all of the desired properties; for example, in multidimensional Lagrangian gas dynamics, it is not possible to construct a discretization that simultaneously conserves energy and preserves entropy in smooth isentropic flows.

Conservation is not the only important property to mimic. Another feature of operators, which is closely related to physics, is the associated null space. In the continuum, the gradient of a scalar function can be zero if and only if this function is constant in space; we say that the null space of the gradient operator consists of constants. Similarly, the null space of the analytic divergence operator consists of vectors that can be represented as a curl of another vector field. If the discrete operators have a larger null space than their continuum counterparts, parasitic (that is, unphysical) modes may grow and pollute the numerical solution. For example, in electromagnetics one may see magnetic monopoles (see discussion in Hyman and Shashkov 1999a). In Lagrangian gas dynamics on a two-

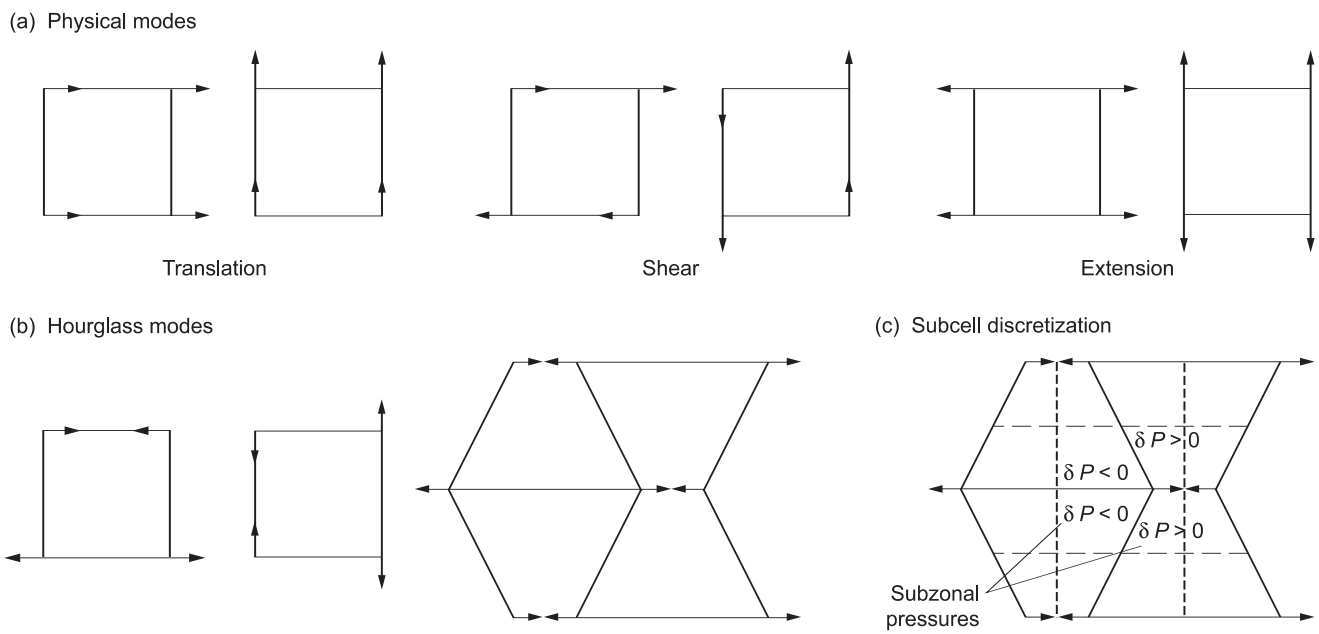


Figure 1. Hourglass Modes
Degrees of freedom that are exhibited by a quadrilateral cell in a Lagrangian mesh are shown in (a) and (b). In addition to physical patterns of motion-translation, extension, shear and rotation, a quadrilateral cell in a Lagrangian mesh can exhibit an unphysical motion called an hourglassing. Because hourglassing neither changes the area of the cell nor does any work on the cell, this pattern produces no restoring forces. Thus, an additional mechanism must be introduced to control the resulting artificial grid distortion. One approach (Margolin and Pyun 1987) to treating hourglassing is to directly filter the pattern from the velocity field. An alternate strategy (Caramana and Shashkov 1998) is to employ a subcell discretization for density (see the dotted lines in Figure 1(c)) that recognizes the consequent hourglass distortion and produces restoring forces (δP in Figure 1(c)).

dimensional (2-D) quadrilateral mesh, one may see so-called hourglassing modes, which distort the shape of the cells without producing restoring forces (refer to Figure 1). This problem is well known in the finite-element community, where it is termed "under-integration;" however, hourglassing patterns are found in finite-difference and finite-volume simulations as well. On the other hand, when the discrete operators have a smaller null space, the solution becomes "stiff," a problem analogous to the well-known phenomenon of locking in finite elements.

The finite size of computational cells leads to another important consideration for mimetic algorithms. While the PDEs can resolve all the scales of motion in a problem, a simulation is more restricted. For example, in high Reynolds number flows, the energy dissipation by molecular viscosity cannot be resolved. The absence of the effects of physical viscosity leads to the need for an artificial mechanism to dissipate a correct amount of energy; in turbulence, this mechanism is called a subgrid-scale model, while in compressible flows with shocks, it is termed an artificial viscosity. Artificial viscosity was first proposed by von Neumann and Richtmyer (1950) to regularize shocks that can not be resolved on the computational mesh. By "regularize," we mean dissipate sufficient energy (and create sufficient entropy) to capture the shock on the mesh without unphysical oscillations. In fluids and gases, the forces due to physical viscosity are isotropic. However, to effectively regularize shocks so that the flow does not depend on the details of the computational mesh, the artificial viscosity needs to have the form of a (possibly nonsymmetric) second-order tensor (Campbell and Shashkov 2001).

In Figure 2, we demonstrate the extent to which a numerical solution can be affected by the choice of mesh if the artificial viscosity is not properly formulated. The simulated problem is known as the Noh implosion and is widely used to study the effects of artificial viscosity. Initial conditions for this problem are specified as a spatially uniform density and an inward radial velocity. The flow has a simple analytic solution, which is an expanding circular shock wave. For the values of density and velocity specified, the position of the shock is at radius

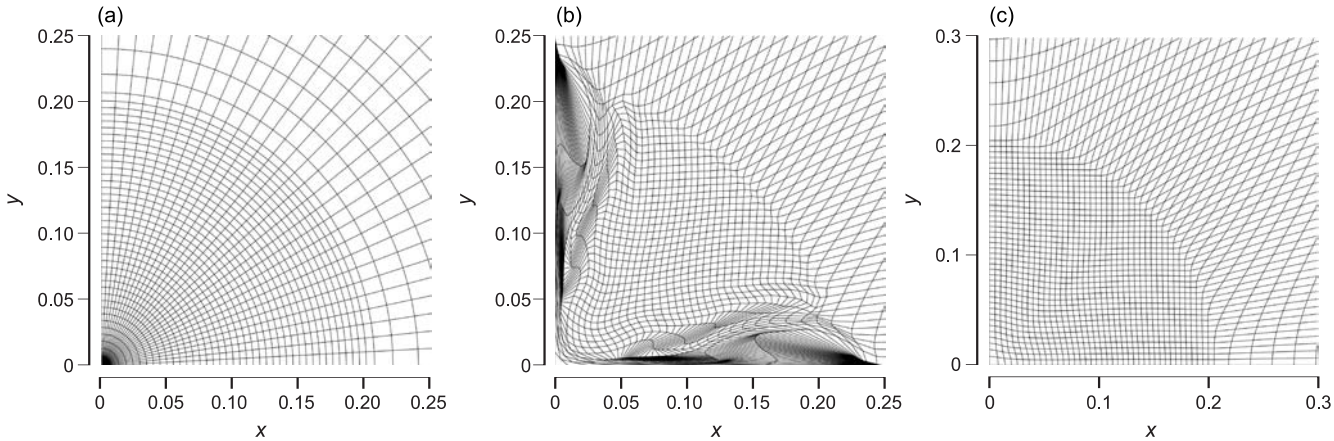


Figure 2. The Effects of the Choice of Mesh and of Artificial Viscosity on an Implosion Problem

(a) Results using an initial mesh with polar symmetry that anticipates the converging fluid flow of a radial implosion. The simulation employs an edge artificial viscosity. The figure shows the simulation is in excellent agreement with the analytic solution at time $t = 0.6$. (b) Results using the same edge viscosity as in panel (a) but starting from a square mesh produces asymmetric results by $t = 0.6$. (c) Results using a tensor artificial velocity and starting with an initially square mesh produces superior results at $t = 0.6$.

$R = 0.2$ for time $t = 0.6$. Now, in Lagrangian simulations, best results are typically obtained when the symmetry of the flow coincides with the symmetry of the mesh. Unfortunately, in realistic problems such a choice of mesh is not always possible. To illustrate these points, we present results for two types of initial mesh: A polar mesh that reflects the anticipated symmetry of the flow is shown in Figure 2(a), and a uniform square mesh is shown in Figures 2(b) and 2(c). Two types of artificial viscosity are used, an "edge viscosity" (Caramana et al. 1998a), as illustrated in Figures 2(a) and 2(b), and a tensor viscosity (Campbell and Shashkov 2001), shown in Figure 2(c). The edge artificial viscosity works well for the initial polar mesh, which is aligned with flow—see Figure 2(a)—but performs poorly for the initial square mesh shown in Figure 2(b), which is not aligned with flow. The reason for such behavior is that the forces generated by the artificial edge viscosity depend strongly on mesh. The tensor artificial viscosity is based on a mimetic discretization of the gradient of a velocity. Because this gradient is based on the discretization of a coordinate invariant differential operator, it is able to produce results that show essentially no dependence on the mesh—see Figure 2(c).

The preservation of the physical flow symmetry in an implosion is critically important to achieve accurate predictions for the inertial confinement fusion program. Small departures from spherical symmetry due to discrete errors can grow into unacceptably large asymmetries in systems undergoing strong convergence. Also, the uncertainty of whether a nonsymmetric result is due to numerical errors or to the physical design severely limits our predictive capability and ultimately our understanding of the dynamical behavior of an implosion. Those methods that preserve symmetries are viable for investigating perturbations of these symmetries. However, the development of such methods may require consideration of meshes with curvilinear edges (as opposed to straight line segments) and the derivation of discrete operators on such a mesh (Margolin and Shashkov 1999, Margolin et al. 2000b). An alternative approach on a line segment mesh has been developed based on the addition of special corrective forces (Caramana and Whalen 1998).

We demonstrate the importance of using symmetry-preserving discretizations on a spherical version of the Rayleigh-Taylor instability (Margolin et al. 2000b). Radial gravity is assumed to act on an unstable interface placed at radius $R = 1$. The computational domain is $.25 \leq R \leq 1.75$. We use a γ-law gas as the equation of state, with $\gamma = 1.4$. The initial velocity for all nodes is zero. The density is 100.0 for $R > 1$ and 1.0 for $R < 1$. The initial pressures are chosen to be in exact hydrostatic balance. The gravitational constant is taken as 0.02.

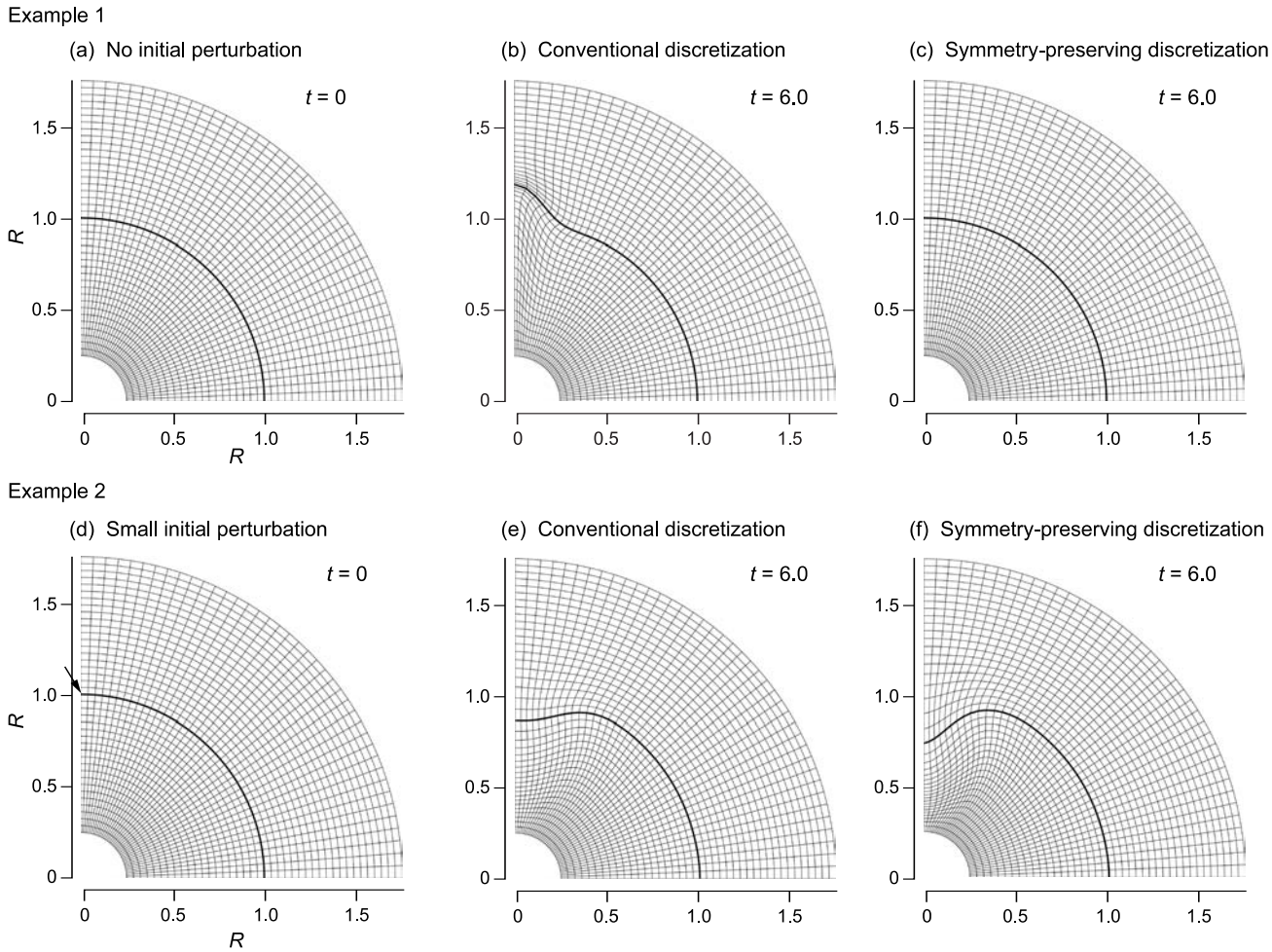


Figure 3. The Effects of Symmetry-Preserving Discretization in the Simulation of a Spherical Rayleigh-Taylor Instability

(a) An interface at $R = 1$ initially separates a dense outer fluid from a less-dense inner fluid. Both fluids are in a gravitational field directed radially inward. (b) With no initial perturbation at the fluid interface, the solution obtained by differencing on a line-segment mesh (Caramana et al. 1998c) develops an unphysical instability by $t = 6.0$. (c) With the same initial conditions, the solution at $t = 6.0$ obtained by differencing on a curvilinear mesh (Margolin and Shashkov 1999) is unchanged from that at $t = 0$ (as expected). (d) The initial mesh is slightly perturbed at the north pole. (e) The solution obtained by differencing on a line-segment mesh shows an instability whose maximum growth rate is not along the vertical axis at $t = 6.0$, which is incorrect. (f) The solution obtained using a curvilinear mesh and with the same initial perturbation as in panel (d) produces an instability whose maximum growth rate is along the vertical axis at $t = 6.0$, which is qualitatively correct.

In the first example shown in Figure 3, the initial state—refer to Figure 3(a)—is represented on a polar grid without any initial perturbation. When a conventional discretization scheme is used (Caramana et al. 1998b), an asymmetric truncation error quickly triggers an instability and, by time $t = 6$, has produced the unphysical mode shown in Figure 3(b). The symmetry-preserving scheme does not trigger any instability. Therefore, the solution at $t = 6$ shown in Figure 3(c) is unchanged from the initial condition in Figure 3(a).

The second example employs a grid with a very small initial perturbation (not visible to the naked eye) at $R = 1$—see Figure 3(d). Let θ_i be the usual angle in

the $r - z$ plane of the points along $R = 1$ in the unperturbed grid. The perturbed grid replaces these points with $r_i = (1 + \mathrm{f}(\theta_i)) \cos(\theta_i)$ and $z_i = (1 + f(\theta_i)) \sin(\theta_i)$. The perturbation f is designed to produce a very small indentation centered at the north pole. The exact form is given by

$$f = -.002\left(1 + 6.5\left(\frac{\pi}{2} - \theta_i\right)^2\right)^{-1} \quad . \tag{4}$$

The solution that uses a conventional scheme is shown in Figure 3(e). It is visibly different from that produced by the symmetry-preserving scheme. The maximum growth rate for the conventional scheme is no longer along the z-axis, even though the initial perturbation is largest at the z-axis.

The solution at time $t = 6$ for this case, using mimetic differencing on curvilinear mesh, is shown in Figure 3(f). It exhibits the expected growth of the initial perturbation. The maximum growth rate is along the z-axis, where the initial perturbation is largest.

As previously noted, the construction of discrete operators and the overall properties of discrete algorithms depend significantly on the choice of the computational mesh. In addition to trying to coordinate the mesh symmetry with the expected symmetry of the flow, it is found that aligning the mesh with material interfaces (Hyman et al. 2002) and having orthogonality of the mesh lines to the interface (Khamayseh and Hansen 2000) are also key to improving the accuracy of simulations. Further, the overall accuracy of an algorithm also depends on the smoothness of the mesh. (A mesh is smooth if such characteristics as the volumes of the cells and the lengths of the cell edges vary smoothly in the mesh—refer to Knupp et al. 2002.)

In Lagrangian simulations, there is no guarantee that an initially smooth mesh will remain smooth. For this reason, a hybrid technique named arbitrary Lagrangian-Eulerian, or ALE, has been developed (Margolin 1997) to allow the automatic identification and improvement of Lagrangian meshes during the simulation. ALE techniques require a strategy for how to rezone (that is, improve) a nonsmooth or tangled mesh. Some elements of this strategy are to preserve the integrity of interfaces and other physically important surfaces (Garimella et al. 2004) and to try to "mass match," that is, to make the mass of the cells vary smoothly in space. However, formulating more general and more complete strategies for rezoning, which simultaneously improve mesh quality while enhancing solution accuracy, is an active field of research.

There are many other issues to consider in the design of discrete operators. For example, for the implicit discretization of a diffusion equation, one needs to solve a system of linear (or perhaps nonlinear) equations. The continuum diffusion operator is symmetric and positive-definite (SPD). If the discrete gradient and divergence are negatively adjoint to each other, then the discrete diffusion operator is also SPD (Hyman et al. 2002). Such SPD operators have the practical advantage that there exist efficient iterative solvers for the associated matrix equations.

To summarize, we have illustrated that many of the important properties of the PDEs that describe the evolution of physical processes are inherent in the differential operators from which they are constructed. We have given examples of how to design discrete operators that mimic these important properties of their analytic counterparts. In some cases, these properties transcend the individual discrete operators and require relationships between different operators to be enforced. We offer that our approach of a discrete tensor and vector analysis pro-

vides a formal framework to study the convergence, symmetries, and accuracy of numerical methods (Berndt et al. 2001). At the same time, we recognize that this is an unfinished story and much work remains to be done.

Balanced Approximations for Time Integration of Multiple-Time-Scale Systems

It can be quite a challenge to do numerical modeling of physical systems that involve many processes occurring at different speeds. The faster processes must be resolved by small simulation time steps, which is computationally expensive, or must be modeled by other means.

Often, the faster processes are nearly in balance at all times, and the system as a whole evolves more slowly than any of the faster processes. A classic example of this type of situation is the flame speed of a laminar diffusion flame. The diffusion and reaction at the flame front are fast processes. However, they compete with each other, with one process slightly dominating the other. The two processes are nearly in balance, producing a flame front that propagates relatively slowly. This is the type of multiple-time-scale problem considered here. There are many examples of such problems in plasma physics, geophysical fluid dynamics, combustion, and radiation hydrodynamics (see, for example, Brackbill and Cohen 1985).

For these problems, it is computationally efficient to resolve only the relatively slow evolution of the system as a whole by using a time step that is large compared with the time scales of the faster processes. At the same time, one must preserve the dynamical balance responsible for the slow evolution of the system. An effective way to achieve this result is to design nonlinear, implicit time-integration schemes that ensure a consistent solution of the separate processes even when large time steps are used. We call such techniques implicitly balanced (Knoll et al. 2003). These techniques were avoided in the past because of a lack of efficient implicit solvers. At that time, formulations based on time splitting and/or linearization were mainly used (Brackbill and Cohen 1985).

In this article, we demonstrate that (1) split methods contain inherent errors that could be dangerous for predictive simulation, (2) modified equation analysis (MEA) (Hirt 1968, Warming and Hyett 1974) can identify possible errors in split methods, and (3) modern, implicitly balanced methods can provide efficient alternatives to split methods. The second point is important because some form of time splitting is required for many problems of interest. We demonstrate these three points by using simple numerical experiments and numerical analysis.

First, we show how MEA can identify splitting errors. The classical analysis of splitting and linearization errors uses asymptotic expansions of exponential operators (Strang 1968). The technique is well suited to determining the stability and assessing the order of accuracy (that is, the rate of convergence) of time-split algorithms. However, the analysis is less useful for obtaining quantitative estimates of the consequences of linearization, the effects of boundary conditions, or the error itself. The latter items can be more readily obtained using MEA, in which a Taylor-series truncation analysis is applied to the discretized PDE (or semidiscretized PDE, for the example considered here). The continuum PDE is reassembled on the left side of the equation, and all the other terms are brought to the right side. This is the new, or modified, equation used for MEA.

Let us now define an implicitly balanced method and compare it with a time-split method, using the equation for the time-dependent reaction-diffusion problem

$$\frac{d\mathbf{u}}{dt} = \mathcal{D}_{\mathbf{u}}\mathbf{u} + \mathcal{R}_{\mathbf{u}}\mathbf{u} \quad , \tag{5}$$

where $\mathbf{u}$ is the dependent variable (or perhaps a system of dependent variables), t is time, $\mathcal{D}_{\mathbf{u}}$ represents the spatial discretization of a diffusion term, and $\mathcal{R}_{\mathbf{u}}$ represents the volumetric reaction, with both $\mathcal{D}_{\mathbf{u}}$ and $\mathcal{R}_{\mathbf{u}}$ being functions of $\mathbf{u}$. In an implicitly balanced method, $\mathcal{R}_{\mathbf{u}}\mathbf{u}$ and $\mathcal{D}_{\mathbf{u}}\mathbf{u}$ will be evaluated at the same value of $\mathbf{u}$ when advancing $\mathbf{u}$ in time. This evaluation is not done with a linearized time-split method.

We wish to advance the solution one discrete time step from the existing time level $\mathbf{u}^n$ to the new time level $\mathbf{u}^{n+1}$. A standard first-order linearized time-split method advances the solution using two linearized subsystems:

$$\frac{\mathbf{u}^* - \mathbf{u}^n}{\Delta t} = \mathcal{D}_{\mathbf{u}}^n \mathbf{u}^* \tag{6}$$

and

$$\frac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^*}{\Delta t} = \mathcal{R}_{\mathbf{u}}^n \tilde{\mathbf{u}}^{n+1} \quad , \tag{7}$$

where $\mathbf{u}^*$ is an intermediate, or temporary, value for $\mathbf{u}$. The effective time step is then given by

$$\frac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^n}{\Delta t} = \mathcal{D}_{\mathbf{u}}^n \mathbf{u}^* + \mathcal{R}_{\mathbf{u}}^n \tilde{\mathbf{u}}^{n+1} \quad , \tag{8}$$

The linearization that has occurred here is in evaluating $\mathcal{D}_{\mathbf{u}}$ and $\mathcal{R}_{\mathbf{u}}$ at the known values of $\mathbf{u}$, $\mathbf{u}^n$.

One possible second-order-accurate implicitly balanced approach would be

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} = \mathcal{D}_{\mathbf{u}}^{n+\frac{1}{2}} \mathbf{u}^{n+\frac{1}{2}} + \mathcal{R}_{\mathbf{u}}^{n+\frac{1}{2}} \mathbf{u}^{n+\frac{1}{2}} \quad . \tag{9}$$

The solution of this time discretization will require a nonlinear iteration involving both diffusion and reaction. It is clear that given the same initial value, $\mathbf{u}^n$, these two methods do not give the same final value at the new time level, that is, $\tilde{\mathbf{u}}^{n+1} \neq \mathbf{u}^{n+1}$. We need to understand when this difference is important for predictive simulation.

We will compare and contrast implicitly balanced methods with a simple linearized time-split method using numerical analysis and numerical experiments with a simple model problem. For further details on this discussion, refer to Knoll et al. (2003). In the following paragraphs, we touch only on issues related to splitting, not on those related to linearization.

We consider only the simplest first-order splitting to illustrate the important points. It is straightforward to design a second-order-accurate splitting for the problem considered below. MEA analysis of more sophisticated splittings is ongoing.

We consider the linear reaction-diffusion problem with T as the scalar dependent variable, a constant diffusivity D, and a constant reactivity $\alpha < 0$:

$$\frac{\partial T}{\partial t} - D\frac{\partial^2 T}{\partial x^2} = \alpha T \quad , \tag{10}$$

with standard boundary conditions and initial conditions. The dynamical time scale is estimated to be

$$\frac{1}{\tau_{\text{dyn}}} \equiv \left|\frac{1}{T}\frac{dT}{dt}\right| \approx \frac{1}{\tau_{\text{dif}}} + \frac{1}{\tau_{\text{reac}}} \quad , \tag{11}$$

where the diffusion time τ_{dif} and reaction time τ_{reac} are

$$\tau_{\text{dif}} \equiv \frac{L^2}{D}; \tau_{\text{reac}} \equiv \left|\frac{1}{\alpha}\right| \quad ,$$

and L is the gradient scale of the solution.

To solve Equation (10), we consider a first-order time-split method that first advances the reaction and then the diffusion. Specifically, the first-order splitting is

$$\frac{T^* - T^n}{\Delta t} = \alpha T^* \quad , \text{ and}$$

$$\frac{T^{n+1} - T^*}{\Delta t} - D\left(\frac{\partial^2 T^{n+1}}{\partial x^2}\right) = 0 \quad , \tag{12}$$

where T^* is an intermediate value for T.

We also consider two balanced methods: one first- and the other second-order accurate. The first-order accurate balanced method is

$$\frac{T^{n+1} - T^n}{\Delta t} - D\left(\frac{\partial^2 T^{n+1}}{\partial x^2}\right) = \alpha T^{n+1} \quad . \tag{13}$$

The second-order accurate balanced method is

$$\frac{T^{n+1} - T^n}{\Delta t} - D\left(\frac{\partial^2 T^{n+\frac{1}{2}}}{\partial x^2}\right) = \alpha T^{n+\frac{1}{2}} \quad , \tag{14}$$

where the intermediate time is defined as,

$$T^{n+\frac{1}{2}} \equiv \frac{T^{n+1} + T^{n}}{2} \quad . \tag{15}$$

Considering the semidiscrete problem in time (that is, ignoring the spatial discretization), we require the Taylor series expansion of T^n in terms of T^{n+1}:

$$T^{n} = T^{n+1} - \Delta t T_t + \frac{\Delta t^2}{2} T_{tt} - \ldots \quad , \tag{16}$$

where $T_t = \partial T/\partial t$. It is straightforward to show that the modified equation for the first-order accurate balanced method is

$$\left[T_t - D\left(\frac{\partial^2 T}{\partial x^2}\right) - \alpha T \right] = \frac{\Delta t}{2} T_{tt} + O\left(\Delta t^2\right) \tag{17}$$

and for the second-order accurate balanced method is

$$\left[T_t - D\left(\frac{\partial^2 T}{\partial x^2}\right) - \alpha T \right] = \frac{\Delta t^2}{24} T_{ttt} + O\left(\Delta t^3\right) \quad . \tag{18}$$

MEA tells us that, when Equation (10) is numerically integrated in time using Equation (13), one is really solving Equation (17). Defining the modified equation for the split method is more subtle.

After the two steps from the split method in Equation (12) have been combined, the effective time step is given by

$$\frac{T^{n+1} - T^{n}}{\Delta t} - D\left(\frac{\partial^2 T^{n+1}}{\partial x^2}\right) = \alpha T^{*} \quad . \tag{19}$$

To perform the MEA, we must eliminate T^n and T^* in favor of T^{n+1} and its time derivatives. As we have seen, T^n can be eliminated using standard Taylor-series expansion. Rather than attempting to write a similar Taylor series for T^*, we can use the second step in the split method itself:

$$T^{*} = T^{n+1} - \Delta t D\left(\frac{\partial^2 T^{n+1}}{\partial x^2}\right) \quad . \tag{20}$$

The modified equation for the splitting method can now be written as

$$\left[T_t - D\left(\frac{\partial^2 T}{\partial x^2}\right) - \alpha T \right] = \frac{\Delta t}{2} T_{tt} \underbrace{- \Delta t \alpha D\left(\frac{\partial^2 T}{\partial x^2}\right)}_{\text{New truncation term}} + O\left(\Delta t^2\right) \quad . \tag{21}$$

Compared with the first-order balanced method, namely, Equation (17), a new first-order truncation term has appeared in the split modified equation. This new

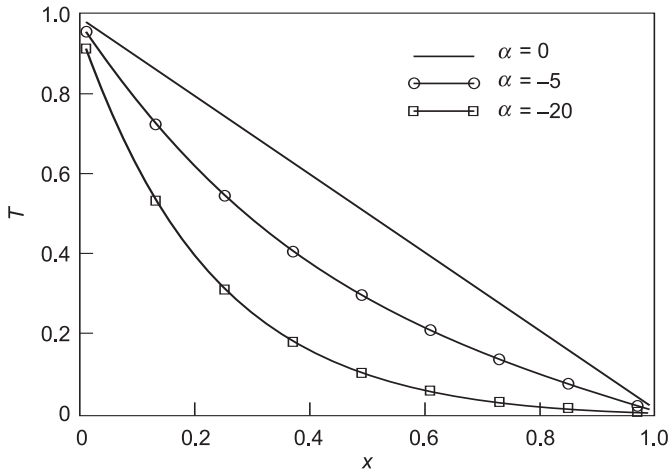


Figure 4. Steady-State Solutions Obtained with the Implicitly Balanced Method
These solutions of the linear reaction-diffusion equation—refer to Equation (10)—were obtained with a second-order-accurate implicitly balanced method. *T* is the scalar dependent variable, and *x* is position. Shown are the solutions for three values of constant reactivity *α*.

term is proportional to the second spatial derivative and scales with $\alpha\,\Delta t$. If an altered diffusion coefficient is used, the modified equation of the split method can be viewed as having the same form as the modified equation of the balanced first-order-accurate method. Indeed, if we replace D with D^* in Equation (21) and equate terms with Equation (17), the result is

$$D^* = \frac{D}{1.0 - \Delta t\alpha} \quad . \tag{22}$$

This suggests that using the split algorithm with the diffusion coefficient D* should reproduce the results of using the first-order-accurate balanced method with the original diffusion coefficient D. For $\alpha < 0$, the altered diffusion coefficient remains positive and less than the original coefficient.

We consider the problem on the domain $0 < x < 1$ with initial conditions $T(x, t = 0) = 0.1$, $T(x = 0, t) = 1$, $T(x = 1, t) = 0.1$, $D = 1$, $\alpha = -20$, and a time step, Δt, of 0.01. To demonstrate some properties of the solution, we have simulated the problem using the second-order-accurate balanced method with $\Delta t = 0.0001$ and $\alpha = -0, -5$, and -20. Figure 4 shows how different values of the finite reaction term α affect the steady-state solutions. Figure 5 shows the time-dependent solutions at $x = 0.1$. At early times, the dynamical time scale is dominated by the diffusion time scale, τ_{dif}, since L is very small near $x = 0$ (the initial gradient is sharp). As this initial structure fades, the impact of finite α on the evolution of the solution becomes clear.

A study of the time-step convergence, verifying that the simple split method is indeed first-order accurate, is given in Knoll et al. (2003). However, it is not apparent from this study that the split method will give the correct steady-state solution using a large time step—that is, $\alpha\Delta t \approx \mathcal{O}(1)$. Figure 6 shows the solutions as functions of time at a particular point ($x = 0.1$) for the different solution methods. For a time step chosen so that $\alpha\Delta t = 0.2$, the split method does not give the correct steady-state solution. The solution from the split method gives no indication of error since the method is stable and qualitatively correct.

In Figure 7, we show the time history of the solution at the same point ($x = 0.1$) for the first-order balanced method and for the split method with the modified diffusion coefficient D^* given in Equation (22). These two solutions are identical, confirming the validity of the MEA of the splitting errors. From these results, it is evident that the solutions given by these first-order split methods can be interpreted as solutions from a balanced method using an altered diffusion

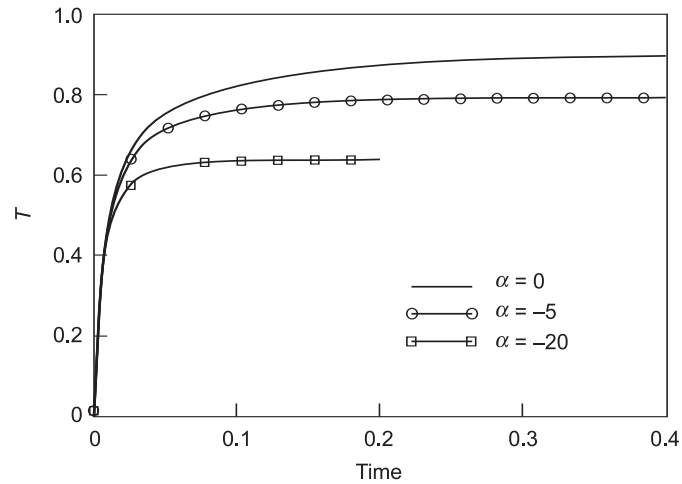


Figure 5. Time-Dependent Solutions Obtained with the Implicitly Balanced Method
These time-dependent solutions to Equation (10) at $x = 0.1$ correspond to the steady-state solutions shown in Figure 4.

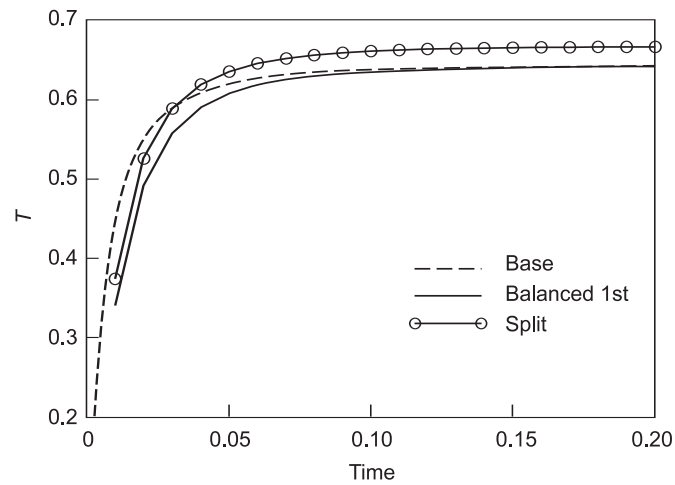


Figure 6. Implicitly Balanced Solutions vs a Split Solution
We compare the time-dependent solutions to Equation (10) at $x = 0.1$ using a second-order-accurate implicitly balanced method ("Base"), a first-order-accurate implicitly balanced method ("Balanced 1st"), and a split method ("Split").

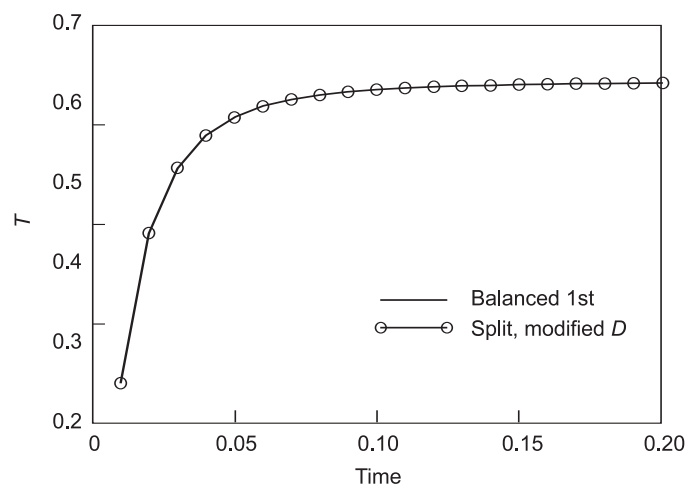


Figure 7. Equivalence of the First-Order-Accurate Balanced Solution and the Split Solution
The solution to Equation (10) obtained with the first-order-accurate balanced method and the original diffusion coefficient ("Balanced, 1st") is identical to the solution obtained with the split method and the corrected diffusion coefficient ("Split, modified *D*"). The original and corrected diffusion coefficients are related through Equation (22).

coefficient. The degree to which the diffusion coefficient is altered is proportional to the chosen time step normalized by a normal mode (fast) time scale, that is, $\alpha\Delta t = \Delta t/\tau_{\text{reac}}$.

Developing implicitly balanced methods that can be used to simulate large

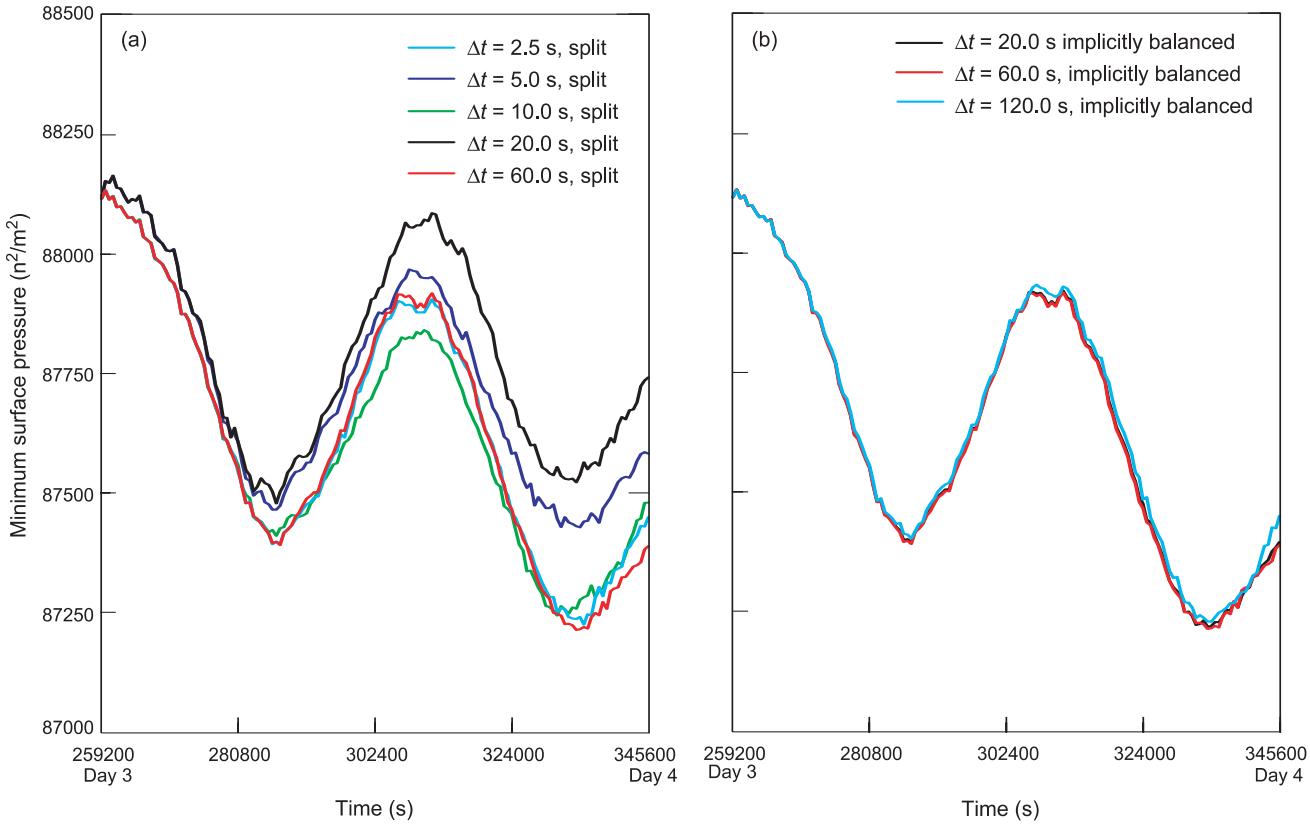


Figure 8. Solution Convergences for the Balanced and Split Methods
We used a 3-D simulation of the minimum pressure in the eye of a hurricane to compare the convergences of (a) split-linearized solutions for different values of the time step (Δt) and (b) implicitly balanced solutions.

three-dimensional (3-D) multiphysics problems is an ongoing research effort that involves many contributors. To give one example, work in this area is discussed in Knoll and Keyes (2004).

Another way to remove the splitting errors is by iterating on the splitting methods. Although some 3-D multiphysics problems have been simulated with implicitly balanced methods, time splitting and linearization are still required for many problems. Thus, we must gain a deeper understanding of the inherent error in time splitting and linearization to achieve more accurate simulations.

Finally, we present results from research using implicitly balanced methods to simulate hurricane intensification (Mousseau et al. 2002, Reisner et al. 2003, Reisner et al. 2004). This 3-D work involves the simulation of compressible multiphase flow. Hurricanes intensify by passing over warm water, and the signature of intensification is the minimum pressure in the hurricane eye. In Reisner et al. (2004), an initially steady-state hurricane is driven into a transient state by specific time-dependent boundary conditions, namely, a time-varying temperature at the ocean's surface. The dynamical time scale in this problem is estimated to be roughly 100 seconds, whereas the sound-wave time scale is roughly 1 second. The split-linearized method, therefore, is used on sound-wave physics equations. Figure 8 shows that, for the implicitly balanced method, the correct solution converges for a time step of $\Delta t = 60$ seconds, whereas the split-linearized method requires a time step of $\Delta t = 1$ second to achieve convergence. In this article, the implicitly balanced method achieved convergence about 5 times faster than the split-linearized method.

Asymptotic-Preserving Discretization Schemes

Asymptotic limits associated with PDEs are limits in which certain terms in an equation are purposely made "small" relative to other terms. Such limits reflect physical situations in which certain physical quantities or processes do, in fact, dominate others. For instance, the compressible hydrodynamic Euler equations, which describe inviscid fluid flow, represent an asymptotic limit of the nonlinear Boltzmann equation for rarefied gas dynamics. In that limit, the ratio of the mean distance between atomic collisions to the system size goes to zero. Similarly, the equations for incompressible fluid flow can be derived from the compressible Euler equations in the limit as the ratio of the material speed to the speed of sound in the material goes to zero. Although asymptotic equations approximate the equations from which they are derived, they accurately represent system behaviors for problems that are highly asymptotic.

An asymptotic equation emerges from the process for obtaining a formal asymptotic solution. The mathematical procedure for obtaining such a solution introduces an asymptotic dimensionless scaling parameter ε that tends to zero. First, the original, or parent, equation is put in dimensionless form, and some of the terms in the equation are scaled by ε^n, where n is a positive integer that may take on different values for different terms. This scaling is defined so that the equation has the desired asymptotic physical behavior as ε goes to zero. Once the scaling is completed, the equation is returned to dimensional form, and the asymptotic solution is assumed to take the form of a power series expansion in ε. This expansion is substituted into the scaled equation, and coefficients of like powers of ε are equated, thereby forming a hierarchical set of equations for the expansion coefficients. The expansion coefficient associated with the lowest power of ε represents the asymptotic solution, that is, the solution obtained in the limit as ε goes to zero. One can use the hierarchical equations to deduce the equation satisfied by the asymptotic solution and thereby obtain the asymptotic equation.

Because the asymptotic equation is generally simpler than the parent equation, it is easier to solve the asymptotic equation for the problems for which it applies than to solve the parent equation. The applicable problems are those in which the assumed dominance of certain terms occurs to a significant extent. Of course, no real problem is perfectly asymptotic, but the exact limit can be approached as closely as desired. As a problem becomes increasingly asymptotic, the solution of the asymptotic equation approaches the solution of the parent equation. However, many problems that require numerical solutions have spatial regions that change in time from asymptotic to nonasymptotic. In those cases, it is often impractical to solve the parent equation in nonasymptotic regions and the asymptotic equation in asymptotic regions. Thus, one must obtain solutions in both the nonasymptotic and asymptotic regions using a single numerical approximation to the parent equation. For the approximation scheme to be valid, solutions to the discrete equation must converge to the continuum solutions as the mesh size goes to zero in both asymptotic and nonasymptotic regions. The problem is that not all methods of discretizing the parent equation produce solutions that converge appropriately in the asymptotic regions. On the contrary, for some discretization schemes, an accurate asymptotic solution is obtained only if the mesh size h resolves length scales much smaller than those relevant to the asymptotic solution. We call such schemes nonasymptotic preserving. Such schemes are inefficient in highly asymptotic regions because they require an excessively large number of spatial cells. In fact, nonasymptotic-preserving schemes require an infinite number of cells in the limit as a region becomes perfectly asymptotic.

To determine whether a discretization scheme "preserves" the asymptotic limit (that is, converges appropriately to the asymptotic solution), one must perform and analyze an asymptotic expansion for the discrete equation that is completely analogous to the expansion for the continuum equation. In this article, we use a particle transport equation and the asymptotic diffusion limit associated with this equation to illustrate both the continuum and discrete asymptotic methodologies. The asymptotic diffusion limit of particle transport is characterized by negligible particle absorption and a diffusion length that is large relative to the mean free path (or average distance between collisions). We derive the diffusion limit for the continuum transport equation and then apply the asymptotic methodology to two spatially discrete forms of the transport equation. One form is obtained using the diamond discretization scheme, and the other is obtained using the upwind discretization scheme. We show that the diamond scheme is asymptotic preserving and the upwind scheme is not. Finally, we give specific computational examples demonstrating the contrasting behavior of these schemes in highly asymptotic (diffusive) problems.

We focus our discussion on a particle transport equation:

$$\mu \frac{\partial vN}{\partial x} + (\sigma_a + \sigma_s)\, vN = \frac{\sigma_s}{2}\int_{-1}^{+1} vN(x,\mu')\,d\mu' + Q \quad . \tag{23}$$

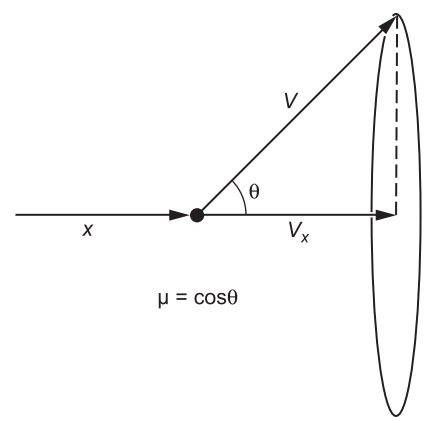


Figure 9. Variable Definitions for the Particle-Density Function *N*(*x*, *μ*)
The number of particles at position *x* moving in direction *μ* is N(x, *μ*) dx d*μ*.

This is an equation for a phase-space particle-density function, $N(x, \mu)$. Although this function depends on a single spatial coordinate, its domain is 3-D and corresponds to an infinite slab. All particles travel at a single speed, v, in directions characterized by the cosine $\mu = v_x/v$. Each cosine corresponds to a cone of directions as illustrated in Figure 9. Particles are assumed to be uniformly distributed within the band. The number of particles located at position x in direction μ, is $N(x, \mu)\, dx\, d\mu$. The spatial volume associated with dx has unit dimensions in the other two Cartesian coordinates, that is, it consists of a differential rectangular box with dimensions $dx \times 1 \times 1$. Particles are randomly absorbed and scattered within the medium. The scattering is isotropic, that is, particles scatter into all directions with equal probability. The absorption cross section is σ_a, and the scattering cross section is σ_s. The expected absorption rate of particles in direction μ at position x is $\sigma_a vN(x, \mu)\, dx\, d\mu$, and the expected scattering rate of particles in direction μ at position x is $\sigma_s vN(x, \mu)\, dx\, d\mu$. The total cross section, σ_t, is the sum of the absorption and scattering cross sections. The mean distance between particle interactions is called the mean free path, and it is given by $\lambda_t = 1/\sigma_t$. The mean free path represents a fundamental spatial scale length in highly absorbing media that appears explicitly in the transport equation. For instance, after traveling a distance s in a purely absorbing medium, a beam of particles is attenuated by a factor of $\exp(-s/\lambda_t)$. The quantity $Q(x,\mu)$ is the particle source function. Therefore, the number of particles created at position x in direction μ is $Q(x,\mu)\, dx\, d\mu$.

Equation (23) is a statement of particle conservation. It simply states that the source rate for the particles entering the differential phase-space volume at position x and direction μ must equal the sink rate for the particles leaving that volume. The boundary conditions for Equation (23) are given in terms of the incident particle distributions at the boundaries. For instance, if the problem domain is the interval [0, 1], the solution to Equation (23) is uniquely determined once N is defined at $x = 0$ for $\mu > 0$ and at $x = 1$ for $\mu < 0$.

It is convenient for our purposes to rewrite Equation (23) as

$$\mu \frac{\partial \psi}{\partial x} + \sigma_t \psi = (\sigma_t - \sigma_a)\phi + Q \quad , \tag{24}$$

where

$$\phi = \frac{1}{2}\int_{-1}^{+1} \psi(x,\mu')d\mu' \quad . \tag{25}$$

The quantity $\psi = vN$ is called the angular flux, and the directional average of ψ, which is denoted by ϕ, is called the scalar flux.

We now begin the derivation of the asymptotic diffusion limit associated with Equation (24). For simplicity, we skip the nondimensionalization process and directly scale Equation (24) by the nondimensional scaling parameter ε:

$$\mu \frac{\partial \psi}{\partial x} + \frac{\sigma_t}{\varepsilon}\psi = \left(\frac{\sigma_t}{\varepsilon} - \varepsilon\sigma_a\right)\phi + \varepsilon Q \quad . \tag{26}$$

Scaling the terms in Equation (24) ensures the following behavior as $\varepsilon \to 0$:

(1) The total cross section scales with ε^{-1} and thus becomes infinite (or, equivalently, the mean free path goes to zero).
(2) The absorption cross section scales with ε and thus goes to zero.
(3) The source scales with ε and thus goes to zero to properly normalize the solution.

Because both the mean distance between collisions and the probability of absorption go to zero, it is not difficult to imagine that the result will be a diffusion process for the particles.

We next assume a power series expansion in ε for the asymptotic solution:

$$\psi = \sum_{n=0}^{\infty} \psi^{(n)}\varepsilon^n \quad . \tag{27}$$

Substituting Equation (27) into Equation (26) and equating coefficients of like powers of ε, we obtain a hierarchical set of equations for the expansion coefficients in Equation (27). After slight algebraic manipulation, the leading-order equation $O(1)$ becomes

$$\psi^{(0)} = \phi^{(0)} \quad . \tag{28}$$

This equation simply states that the leading-order solution is isotropic, that is, independent of direction. After considerable manipulation and use of Equation (28), the $O(\varepsilon)$ equation becomes

$$\psi^{(1)} = -\frac{\mu}{\sigma_t}\frac{\partial \phi^{(0)}}{\partial x} + \phi^{(1)} \quad . \tag{29}$$

The $O(\varepsilon^2)$ equation (after considerable manipulation and use of previous equations) becomes

$$\mu\left(-\frac{\mu}{\sigma_t}\frac{\partial\phi^{(0)}}{\partial x}+\phi^{(1)}\right)+\sigma_t\psi^{(2)}=\sigma_t\phi^{(2)}-\sigma_a\phi^{(0)}+Q \quad . \tag{30}$$

Averaging Equation (30) over all μ (by integration), we find that the leading-order solution, $\psi^{(0)} = \phi^{(0)}$, satisfies the following diffusion equation:

$$-\frac{\partial}{\partial x}\left(\frac{1}{3\sigma_t}\frac{\partial\phi^{(0)}}{\partial x}\right)+\sigma_a\phi^{(0)}=Q \quad . \tag{31}$$

Thus, we see that this asymptotic scaling does indeed lead to a limit in which the transport solution satisfies a diffusion equation. The effective boundary conditions satisfied by the asymptotic diffusion solution must be determined by a boundary-layer analysis beyond the scope of this discussion. It suffices to note that, with no incoming particles at the boundaries, the asymptotic diffusion solution is zero at both boundaries.

The fundamental scale length associated with the diffusion equation is the diffusion length L:

$$L=\frac{1}{\sqrt{3\sigma_a\sigma_t}}=\sqrt{\frac{\lambda_t}{3\sigma_a}} \quad . \tag{32}$$

Homogeneous solutions of Equation (31) have the form $\exp(\pm x/L)$. Note that, if we apply the asymptotic scaling defined in Equation (26) to L, we find that L is independent of ε, which is appropriate because an asymptotic scale length should not depend on ε. Further note that, since L is $O(1)$ and λ_t is $O(\varepsilon)$ in the diffusion limit, the mean free path becomes infinitely small relative to a diffusion length in the asymptotic diffusion limit. This implies that the mean free path can be arbitrarily small relative to a diffusion length in problems that are highly diffusive.

The diffusion limit for a spatially discretized transport equation is completely analogous to that for the analytic transport equation. We have shown that the transport solution satisfies an analytic diffusion equation in the asymptotic diffusion limit. By analogy, one would expect a spatially discrete transport solution to satisfy a valid spatially discrete diffusion equation in the asymptotic diffusion limit. A transport spatial-discretization scheme preserves the asymptotic diffusion limit when this occurs. In a practical sense, this means that an accurate solution can be expected in highly diffusive problems if the width of each mesh cell is small compared with a diffusion length. If a discretization scheme does not preserve the diffusion limit, one generally finds that an accurate solution can be obtained for highly diffusive problems only if the width of each cell is small with respect to a mean free path. This condition is nonphysical in the sense that the mean free path is an appropriate scale length for the transport solution in highly absorbing problems, but it is not a scale length for the transport solution in diffusive problems. More significantly, as a problem becomes increasingly diffusive, the mean free path approaches zero while the diffusion length remains constant. Thus, an arbitrarily large number of spatial cells can be required to obtain an accurate solution in highly diffusive problems if a spatial-discretization scheme does not preserve the asymptotic diffusion limit.

We next consider two spatial-discretization schemes for the transport equation and discuss their properties for diffusive problems. The first is the upwind

scheme, and the second is the diamond scheme. Although it may not be obvious, all transport discretizations are completely defined by the equations for a single spatial cell. The reason is that each spatial cell can be considered to be an independent transport domain with the incoming angular flux defined by either true boundary conditions or the outgoing angular fluxes from adjacent cells. Let us consider a cell defined over the interval $[x_{i-1/2}, x_{i+1/2}]$, and let $h = x_{i+1/2} - x_{i-1/2}$ denote the cell width. Integrating Equation (24) over this interval, we get the balance equation, which is exact:

$$\mu\left(\psi_{i+\frac{1}{2}} - \psi_{i-\frac{1}{2}}\right)\Big/h + \sigma_{\mathrm{t},i}\,\psi_i = (\sigma_{\mathrm{t}} - \sigma_{\mathrm{a}})\,\phi_i + Q_i \quad . \tag{33}$$

For simplicity, we have assumed a uniform grid with constant cross sections in Equation (33). Three angular fluxes appear in this equation, namely, two cell-edge values and one cell-average value. As previously noted, the incoming cell-edge angular flux is known, leaving two unknowns: the cell-average and the outgoing cell-edge angular fluxes. The balance equation provides one of two equations needed to close the system. The second equation is usually called the auxiliary equation and relates the outgoing cell-edge and cell-average angular fluxes. In the case of upwind differencing, the outgoing cell-edge angular flux is equal to the cell-average angular flux:

$$\begin{aligned} \psi_i &= \psi_{i+\frac{1}{2}} \quad \text{for } \mu > 0 \,, \\ &= \psi_{i-\frac{1}{2}} \quad \text{for } \mu < 0 \,. \end{aligned} \tag{34}$$

In the case of diamond differencing, the cell-average angular flux is the arithmetic average of the incoming and outgoing cell-edge angular fluxes:

$$\psi_i = \frac{1}{2}\left(\psi_{i+\frac{1}{2}} + \psi_{i-\frac{1}{2}}\right) \quad \text{for all } \mu \quad . \tag{35}$$

An asymptotic analysis for the upwind scheme in the thick diffusion limit yields a rather bizarre result. In particular, the upwind asymptotic solution satisfies the following difference equation:

$$\frac{1}{4h}\left(\phi_i^{(0)} - \phi_{i-1}^{(0)}\right) - \frac{1}{4h}\left(\phi_i^{(0)} - \phi_{i+1}^{(0)}\right) = 0 \quad . \tag{36}$$

If we multiply Equation (36) by $4/h$, we obtain a standard three-point cell-centered discretization for the following analytic diffusion equation:

$$-\frac{\partial^2 \phi^{(0)}}{\partial x^2} = 0 \quad . \tag{37}$$

However, comparison with Equation (31) shows that this is not the right diffusion equation. It contains no cross sections and no source! Thus, the upwind scheme does not preserve the asymptotic diffusion limit.

An asymptotic analysis of the diamond scheme in the thick diffusion limit indicates that the diamond solution satisfies the following asymptotic difference equation:

$$-\frac{1}{3\sigma_t}\left(\phi^{(0)}_{i+\frac{3}{2}} - 2\phi^{(0)}_{i+\frac{1}{2}} + \phi^{(0)}_{i-\frac{1}{2}}\right)\Big/h^2 + \frac{\sigma_a}{4}\left(\phi^{(0)}_{i+\frac{3}{2}} + 2\phi^{(0)}_{i+\frac{1}{2}} + \phi^{(0)}_{i-\frac{1}{2}}\right) = \frac{1}{2}(Q_{i+1} + Q_i) \quad . \qquad (38)$$

This is a valid discretization scheme for the diffusion equation given in Equation (31). Thus, the diamond scheme preserves the asymptotic diffusion limit.

We next consider computational examples that will hopefully make the concept of the discrete diffusion limit concrete. To illustrate the discrete asymptotic limit, we first define a fixed initial transport problem and associate it with $\varepsilon = 1$. The problem then changes as a function of ε, according to the scaling of the total cross section, the absorption cross section, and the source given in Equation (26). Specifically, the initial problem is defined as follows:

(1) The spatial domain is the interval [0, 1], measured in centimeters, and is fixed for all ε.
(2) The transport solution satisfies vacuum boundary conditions, that is, ψ is zero at both boundaries in the incoming directions.
(3) The internal source is spatially constant with $Q = 1$ particle per cubic centimeter per second [p/(cm^3-s)].
(4) The cross sections are spatially constant with $\sigma_t = 10$ expected interactions per centimeter and $\sigma_a = 0.1$ expected absorption per centimeter.
(5) The cell thickness, h, is 0.1 centimeter, for a total of 10 spatial cells.

As previously stated, we assume that this initial problem corresponds to $\varepsilon = 1$. Then, we scale σ_t by ε^{-1}, σ_a by ε, and Q by ε. For instance, when $\varepsilon = 0.1$, we find that $\sigma_t = 100$ expected interactions per centimeter, $\sigma_a = .01$ expected absorption per centimeter, and $Q = 0.1$ p/(cm^3-s). The asymptotic transport solution to this sequence of problems satisfies Equation (31) with zero Dirichlet boundary conditions; that is, the solution is zero at both boundaries. Note that the diffusion equation is invariant to the scaling of the physical parameters, so the set of physical parameters for any value of ε may be used to evaluate the asymptotic diffusion solution. Furthermore, note that h/λ_t is scaled by ε^{-1}, so the number of mean free paths per cell becomes infinite as $\varepsilon \to 0$.

We plot the upwind and diamond solutions in Figures 10 and 11, respectively, for ε =1, 0.25, and 0.1. Figure 11 shows that the upwind solutions converge to zero with decreasing ε, in accordance with the analysis. This convergence to zero occurs because particles enter the computational domain only through the internal source Q, which is not present in the discrete asymptotic equation given by Equation (14). It can be seen from Figure 3 that the diamond solutions appear to converge to the analytic asymptotic diffusion solution given by Equation (31). However, the convergence will eventually stagnate because the mesh is fixed. The diamond solutions actually converge to the solution of Equation (38) with boundary conditions corresponding to $\phi^{(0)} = 0$ at both $x = 0$ centimeter and $x = 1$ centimeter.

We next demonstrate the excessive mesh refinement required by a scheme that does not preserve the diffusion limit. In particular, we plot the upwind solutions for the problem corresponding to $\varepsilon = 0.1$ calculated with 10, 100, and 1000 spatial

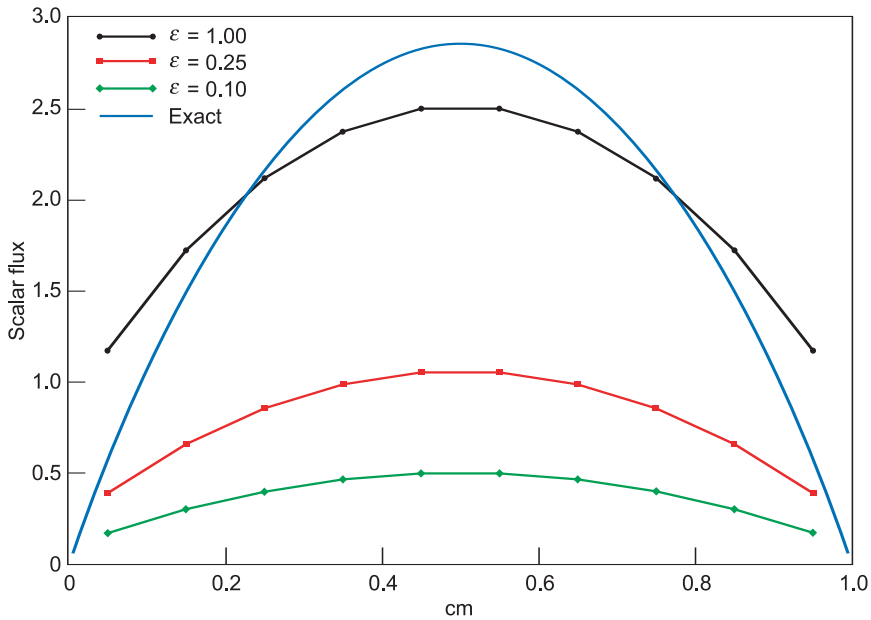


Figure 10. Solutions for the Scalar Flux in the Asymptotic Diffusion Limit with Upwind Spatial Differencing
Shown are the numerical solutions to Equation (31) for several values of ε. The "exact" analytical solution is also shown.

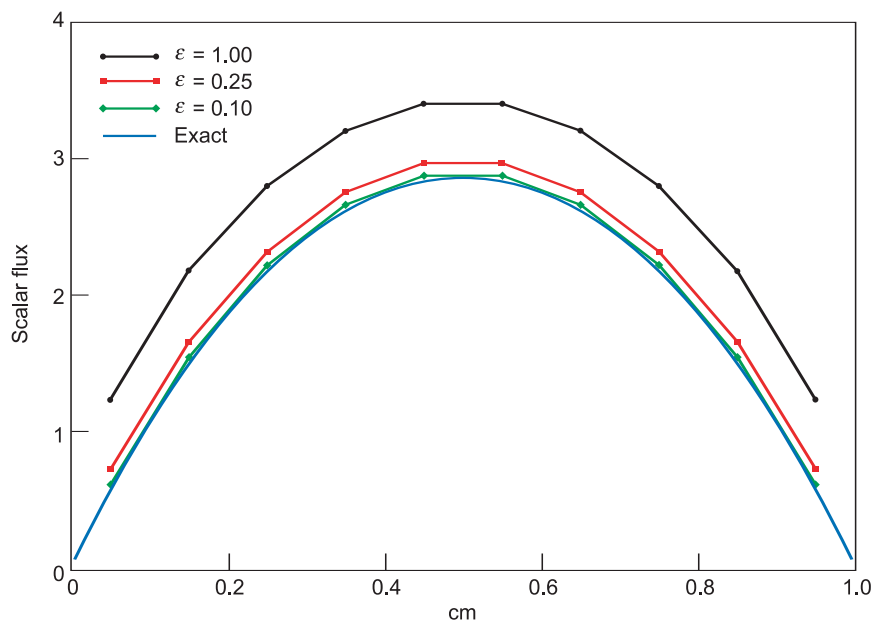


Figure 11. Solutions for the Scalar Flux in the Asymptotic Diffusion Limit with Diamond Spatial Differencing
Shown are the numerical solutions to Equation (31) for several values of ε. The "exact" analytical solution is also shown.

cells, respectively. It can be seen from Figure 12 that the upwind scheme is converging, but a small amount of error is still evident with 1000 spatial cells. The cell thickness in the 1000-cell calculation is 0.01 mean free paths. As expected, an accurate solution requires a cell width that is small when measured in mean free paths. The accuracy of the 1000-cell calculation will be maintained for smaller values of ε only if the cell width remains fixed when measured in mean free paths.

This is why schemes that do not preserve the diffusion limit can require an arbitrarily large number of mesh cells in highly diffusive problems. For instance, one would have to use 10,000 spatial cells for the $\varepsilon = 0.01$ problem to obtain essentially the same solution as with 1000 cells for $\varepsilon = 0.1$. In general, the number of cells required to maintain a given level of accuracy will be inversely proportional to ε. This is to be contrasted with the asymptotic-preserving diamond

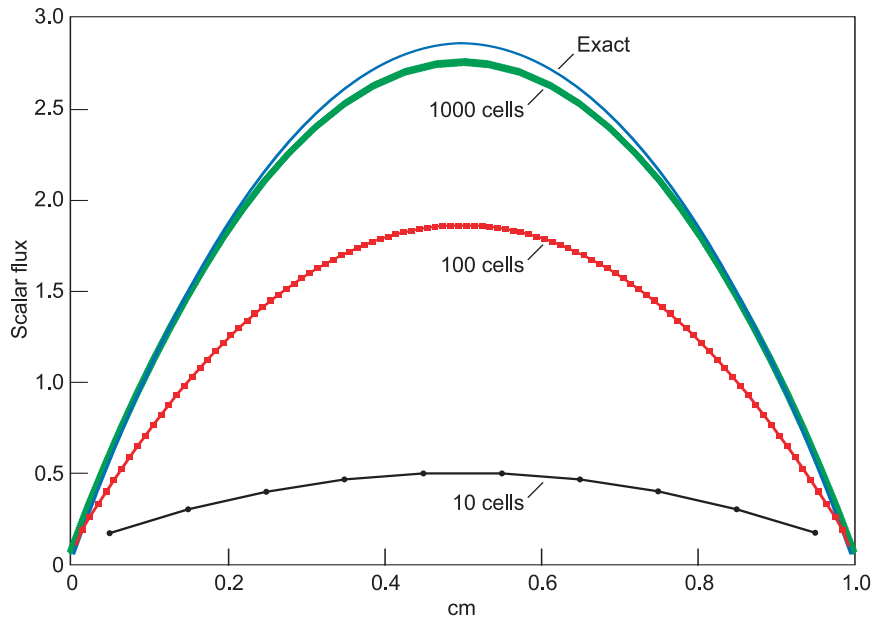


Figure 12. Solutions for the Scalar Flux in the Asymptotic Diffusion Limit with Upwind Differencing
Shown are the numerical solutions to Equation (31) for $\varepsilon = 0.1$ and different numbers of computational cells. The "exact" analytical solution is also shown.

scheme, which maintains a given level of accuracy with a fixed number of cells in the limit as $\varepsilon \to 0$, even though the cell width measured in mean free paths becomes infinite in this limit.

In summary, it is essential to use asymptotic-preserving discretization schemes in asymptotic problems whenever the scale lengths associated with the asymptotic equation are much larger than one or more scale lengths that explicitly appear in the parent equation. Schemes that are not asymptotic preserving can be prohibitively expensive to use because they require the mesh to be refined with respect to scale lengths that can be arbitrarily small compared with the scale lengths associated with the asymptotic solution. Although we have focused on the transport equation and the asymptotic diffusion limit, the basic properties that we have illustrated apply to a wide variety of physical systems. The concept of asymptotic-preserving discretizations is relatively new and not well known in the computational community. However, it can be expected to gain widespread attention in the near future because of the increasing emphasis on multiphysics/multiscale numerical simulation.

Conclusions

The common thread of the three numerical methodologies discussed in this article is the inclusion of physical insight. Perhaps, the major driving force at Los Alamos for developing such methodologies is the weapons program. However, these methods are also affecting such diverse areas as weather simulation, magnetic-confinement fusion simulations, nuclear reactor safety simulation, and aircraft design. Efforts aimed at developing and implementing such methods are ongoing within several Los Alamos programs. However, developing physically motivated numerical-discretization schemes remains a challenging task as we move toward more-accurate computer simulations of phenomena involving many types of physics. ■

Further Reading

Mimetic Discretizations for PDEs

Berndt, M., K. Lipnikov, D. Moulton, and M. Shashkov. 2001. Convergence of Mimetic Finite Difference Discretizations of the Diffusion Equation. East-West *J. Numer. Math.* **9** (4): 265.

Campbell, J. C., and M. J. Shashkov. 2001. A Tensor Artificial Viscosity Using a Mimetic Finite Difference Algorithm. *J. Comput. Phys.* **172** (2): 739.

Campbell, J. C., J. M. Hyman, and M. J. Shashkov. 2002. Mimetic Finite Difference Operators for Second-Order Tensors on Unstructured Grids. *Comput. Math. Appl*. **44** (1–2): 157.

Caramana, E. J., and M. J. Shashkov. 1998. Elimination of Artificial Grid Distortion and Hourglass-Type Motions by Means of Lagrangian Subzonal Masses and Pressures. *J. Comput. Phys*. **142** (2): 521.

Caramana, E. J., and P. P. Whalen. 1998. Numerical Preservation of Symmetry Properties of Continuum Problems. *J. Comput. Phys.* **141** (2): 174.

Caramana, E. J., M. J. Shashkov, and P. P. Whalen. 1998a. Formulations of Artificial Viscosity for Multi-dimensional Shock Wave Computations. *J. Comput. Phys.* **144** (1): 70.

Caramana, E. J., D. E. Burton, M. J. Shashkov, and P. P. Whalen. 1998b. The Construction of Compatible Hydrodynamics Algorithms Utilizing Conservation of Total Energy. *J. Comput. Phys.* **146** (1): 227.

Garimella, R. V., M. J. Shashkov, and P. M. Knupp. 2004. Triangular and Quadrilateral Surface Mesh Quality Optimization Using Local Parameterization. *Comp. Methods Appl. Mech. Eng.* **193**: 913.

Hyman, J. M., and M. Shashkov. 1997a. Adjoint Operators for the Natural Discretizations of the Divergence, Gradient and Curl on Logically Rectangular Grids. *Appl. Numer. Math.* **25** (4): 413.

———. 1997b. Natural Discretizations for the Divergence, Gradient, and Curl on Logically Rectangular Grids. *Comput. Math. Appl*. **33** (4): 81.

Hyman, J. M., and M. Shashkov. 1999a. Mimetic Discretizations for Maxwell's Equations. *J. Comput. Phys.* **151** (2): 881.

———. 1999b. The Orthogonal Decomposition Theorems for Mimetic Finite Difference Methods. *SIAM J. Numer. Anal.* **36** (3): 788.

Hyman, J. M., S. Li, P. Knupp, and M. Shashkov. 2000. An Algorithm for Aligning a Quadrilateral Grid with Internal Boundaries. *J. Comput. Phys.* **163** (1): 133.

Hyman, J., J. Morel, M. Shashkov, and S. Steinberg. 2002. Mimetic Finite Difference Methods for Diffusion Equations. *Comput. Geosci.* **6** (3): 333.

Khamayseh, A., and G. Hansen. 2000. Quasi-Orthogonal Grids with Impedance Matching. *SIAM J. Sci. Comput.* **22** (4): 1220.

Knupp, P., L. Margolin, and M. Shashkov. 2002. Reference Jacobian Optimization-Based Rezone Strategies for Arbitrary Lagrangian Eulerian Methods. *J. Comput. Phys.* **176** (1): 93.

Margolin, L. G. 1997. Introduction to An Arbitrary Lagrangian-Eulerian Computing Method for all Flow Speeds. *J. Comput. Phys.* **135**: 198.

Margolin, L. G., and J. J. Pyun. 1987. A Method for Treating Hourglass Patterns. In *Proceedings of the Fifth International Conference on Numerical Methods in Laminar and Turbulent Flow*. (Montreal, Canada, July 6–10, 1987). Edited by C. Taylor, W. G. Habashi. And M. M. Hafez. U.K.: Pineridge Press.

Margolin, L., and M. Shashkov. 1999. Using a Curvilinear Grid to Construct Symmetry-Preserving Discretizations for Lagrangian Gas Dynamics. *J. Comput. Phys.* **149** (2): 389.

Margolin, L. G., M. Shashkov, and P. K. Smolarkiewicz, 2000a. A Discrete Operator Calculus for Finite Difference Approximations. *Comp. Methods Appl. Mech. Eng.* **187** (3–4): 365.

Margolin, L., M. Shashkov, and M. Taylor. 2000b. Symmetry-Preserving Discretizations for Lagrangian Gas Dynamics. In *Proceedings of the 3rd European Conference on Numerical Mathematics and Advanced Applications*. Edited by P. Neittaanmäki, T. Tiihonen, and P. Tarvainen. p. 725. Singapore: World Scientific.

Shashkov, M. 1996. Conservative Finite-Difference Methods on General Grids. Edited by S. Steinberg. Boca Raton, FL: CRC Press.

Von Neumann, J., and R. D. Richtmyer. 1950. A Method for the Numerical Calculation of Hydrodynamic Shocks. *J. Comput. Phys.* **21** (3): 232.

Balanced Approximations for Time Integration of Multiple-Time-Scale Systems

Brackbill, J. U., and B. I. Cohen. 1985. *Multiple Time Scales*. Orlando: Academic Press.

Hirt, C. W. 1968. Heuristic Stability Theory for Finite-Difference Equations. *J. Comput. Phys.* **2** (4): 339.

Knoll, D. A., and D. E. Keyes. 2004. Jacobian-Free Newton-Krylov Methods: A Survey of Approaches and Applications. *J. Comput. Phys.* **193**: 357.

Knoll, D. A., L. Chacon, L. G. Margolin, and V. A. Mousseau. 2003. On Balanced Approximations for Time Integration of Multiple Time Scale Systems. *J. Comput. Phys.* **185**: 583.

Mousseau, V. A., D. A. Knoll, and J. M. Reisner. 2002. An Implicit Nonlinearly Consistent Method for the Two-Dimensional Shallow-Water Equations with Coriolis Force. *Mon. Weather Rev.* **130** (11): 2611.

Reisner, J., A. Wyszogrodzki, V. Mousseau, and D. A. Knoll. 2003. An Efficient Physics-Based Preconditioner for the Fully Implicit Solution of Small-Scale Thermally Driven Atmospheric Flows. *J. Comput. Phys.* **189**: 30.

Reisner, J., V. Mousseau, A. Wyszogrodzki, and D. A. Knoll. 2004. An Implicitly Balanced Hurricane Model with Physics-Based Preconditioning. (To be published in *Mon. Weather Rev*.)

Strang, G. 1968. On the Construction and Comparison of Difference Schemes. *SIAM J. Numer. Anal.* **5** (3): 506.

Warming, R. F., and B. J. Hyett. 1974. The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods. *J. Comput. Phys*. **14** (2): 159.

Asymptotic-Preserving Discretization Schemes

Adams, M. L. 2001. Discontinuous Finite Element Transport Solutions in Thick Diffusive Problems. *Nucl. Sci. Eng.* **137**: 298.

Adams, M. L., and P. F. Nowak. 1998. Asymptotic Analysis of a Computational Method for Time- and Frequency-Dependent Radiative Transfer. *J. Comput. Phys.* **146** (1): 366.

Larsen, E. W., and J. E. Morel. 1989. Asymptotic Solutions of Numerical Transport Problems in Optically Thick, Diffusive Regimes II. *J. Comput. Phys.* **83** (1): 212.

Larsen, E. W., J. E. Morel, and W. F. Miller Jr. 1987. Asymptotic Solutions of Numerical Transport Problems in Optically Thick, Diffusive Regimes. *J. Comput. Phys.* **69** (2): 283.

Lowrie, R. B., and J. E. Morel. 2002. Methods for Hyperbolic Systems with Stiff Relaxation. *Int. J. Numer. Methods Fluids* **40**: 413.

Morel, J. E., T. A. Wareing, and K. Smith. 1996. A Linear-Discontinuous Spatial Differencing Scheme for Sn Radiative Transfer Calculations. *J. Comput. Phys.* **128** (2): 445.

For further information, contact Dana Knoll (505) 667-7467 (nol@lanl.gov) or Len Margolin (505) 665-1947 (len@lanl.gov).

Erratum to "Photoelectron Spectroscopy of Alpha- and Delta-Plutonium"
Los Alamos Science **26**: 168, 2000
A. J. Arko, J. J. Joyce, L. A. Morales, J. H. Terry, and R. K. Schulze

Some of the plutonium research presented in the article was conducted at the Advanced Light Source (ALS), Lawrence Berkeley National Laboratory. The ALS work was performed as a multi-institutional collaboration. In addition to the authors listed for the ALS work (J. H. Terry and R. K. Schulze), we would like to acknowledge their coworkers, who were Jim Tobin of Lawrence Livermore National Laboratory; Tom Zocco and Doug Farr of Los Alamos National Laboratory; David Shuh, Eli Rotenberg and Keith Heinzelman of Lawrence Berkeley National Laboratory; and Peter Boyd of Boyd Technologies. Further details of this portion of the plutonium research are available through the following publications: J. Terry, R. K. Schulze, J. D. Farr, T. Zocco, K. Heinzelman, E. Rotenberg, D. K. Shuh, G. van der Laan, D. A. Arena, and J. G. Tobin. 2002. 5f Resonant Photoemission from Plutonium. *Surf. Sci. Lett.* **499**: L141; J. G. Tobin, B. W. Chung, R. K. Schulze, J. Terry, J. D. Farr, D. K. Shuh, K. Heinzelman, E. Rotenberg, G. D. Waddill, and G. van der Laan. 2003. Resonant Photoemission in *f*-Electron Systems: Pu and Gd. *Phys. Rev. B* **68**: 155109.

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. All company names, logos, and products mentioned herein are trademarks of their respective companies. Reference to any specific company or product is not to be construed as an endorsement of said company or product by the Regents of the University of California, the United States Government, the US Department of Energy, nor any of their employees. The Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

LA-UR-04-8635

Los Alamos
NATIONAL LABORATORY
EST. 1943

The World's Greatest Science
Protecting America