Title
Workshop on Statistical Approaches for the Evaluation of Complex Computer Models

Authors
Berk, Richard
Bickel, Peter
Campell, Katherine
et al.

Publication Date
2001

DRAFT

LA-UR-00-5034

Approved for public release;
distribution is unlimited.

Title: Workshop on Statistical Approaches for the Evaluation of Complex Computer Models

Author(s): Richard A. Berk, Peter Bickel, Katherine Campbell, Sallie Keller-McNulty, Elizabeth Kelly, and Jerome Sacks,

Contributors: Richard A. Berk, Robert Fovell, Rodman Linn, Frederic Schoenberg, Nagui Rouphail, Jerome Sacks, Byungkyu Park, Alan Perelson and Peter Bickel

Submitted to: Statistical Science

DRAFT

Los Alamos
NATIONAL LABORATORY

Workshop on Statistical Approaches for the Evaluation of Complex Computer Models

Richard A. Berk, Peter Bickel, Katherine Campbell, Sallie Keller-McNulty, Elizabeth Kelly, and Jerome Sacks,
Contributors: Richard A. Berk, Robert Fovell, Rodman Linn, Frederic Schoenberg, Nagui Rouphail, Jerome Sacks, Byungkyu Park, Alan Perelson and Peter Bickel

Abstract. As decision- and policy-makers come to rely increasingly on estimates and simulations produced by computerized models of the world, in areas as diverse and climate prediction, transportation planning, economic policy and civil engineering, the need for objective evaluation of the accuracy and utility of such models likewise becomes more urgent. This article summarizes a two-day workshop that took place in Santa Fe, New Mexico in December 1999, whose focus was the evaluation of complex computer models. Approximately half of the workshop was taken up with formal presentation of four computer models by their creators, each paired with an initial assessment by a statistician. These prepared papers are presented, in shortened form, in Section 3 of this paper. The remainder of the workshop was devoted to introductory and summary comments, short contributed descriptions of related models, and a great deal of floor discussion, which was recorded by assigned rapporteurs. These are presented in Sections 2 and 4 in the paper. In the introductory and concluding sections we attempt to summarize the progress made by the workshop and suggest next steps.

Key words and phrases: model accuracy, model evaluation, model validation, uncertainty analysis, computer experiments, statistically equivalent models, model-based decisions

1. INTRODUCTION

Complex computer models for the simulation of real world systems are used pervasively in scientific research, and there are increasing demands for these models to support policy- and decision-making. The foundations for such models range from the best available scientific theory ("structurally valid models" in the terminology of Zeigler [1976]) to empirical observation, common sense, and computational convenience. Frequently, they are assembled by coupling a number of simpler models. A key question is how good these complex computer models

really are for their intended purposes. A subsequent question is how to make the models better.

Such issues prompted the National Academy of Sciences Committee on Applied and Theoretical Statistics (CATS) to initiate planning for and eventually co-host a workshop on computer model evaluation[1] with Los Alamos National Laboratory and the National Institute of Statistical Sciences. The goal of the workshop was to bring modelers, applied mathematicians, and statisticians together to consider the role statistical concepts and tools could play in these evaluations.

The workshop drew a cross-section of nearly 100 scientists from academia, industry, and government. Participants included modelers from the physical and biological sciences, applied mathematicians, and statisticians, with nearly half the participants being modelers with subject-matter expertise. Participants were asked to focus on four key questions throughout the workshop:

(1) **What do we mean by model evaluation?** While there is no dispute that computer models should usefully approximate the real-world phenomena at issue, there is often lack of clarity about what aspects of the approximation are important within the proposed research or policy context. Moreover, under the rubric of "model evaluation" can be found everything from eyeball assessments to rigorous and formal characterizations, with some judgments derived from an aggregate measure of fit and others from a focus on certain key features of model output.

(2) **What makes model evaluation difficult?** There is a daunting abundance of complications: poor calibration; lack of data as "ground truth;" misalignment of temporal and spatial scales between the model output and available data; costs of both simulation and data replicates; large numbers of free parameters; and variability in modeled phenomena, data, and in the simulation itself.

(3) **What strategies for model evaluation can be employed?** Model evaluations have been undertaken for decades by subject-area experts, and there is, at the very least, extensive lore about how this should be done. An important goal of the workshop was to collect and organize this lore and then suggest other possible model evaluation strategies.

(4) **What is the role of statistical concepts and tools, and where are the statistical gaps?** There are a number of model evaluation concerns that can already be addressed by (computationally feasible) statistical methods. But an important task of the workshop was to systematically reconsider these methods in the context of particular model evaluation strategies: for instance, a method

[1] Workshop on Statistical Approaches for the Evaluation of Complex Computer Models was held December 3-4, 1999.

that may work well for comparing computer model output to data may not work well for comparing two sets of computer model output. A related task was to document serious gaps between what may be needed for model evaluation and what statisticians can currently provide.

The workshop began with a lively keynote address by Dr. William Press, the Deputy Director for Science, Technology, and Programs at Los Alamos National Laboratory (summarized in Section 2), the goal of which was to help frame the issues. This was followed by four sessions, each focused on a model drawn from a different scientific application:

- Meteorology
- Wildfire Control
- Transportation Planning
- Immune System Function

In each session, an overview of the computer model was provided by a subject-area scientist, followed by a statistical presentation discussing model evaluation approaches and problems in the context of the specific application model. These presentations laid the foundation for extensive discussion from the floor. Each session ended with a rapporteur placing that discussion in the context of the four workshop questions. The presentations are in Section 3; the discussions and rapporteur syntheses are presented in Section 4. Section 5 revisits the central themes of the workshop and provides a more complete discussion of important questions identified during the workshop (questions needing statisticians' attention), and offers suggestions about how some initial progress might be made. The material to follow is wide ranging. In the interest of providing a structure on which to hang the many details to follow, we offer immediately below the general "sense of the body" that we extracted from the workshop.

1.1 Workshop Impact and Themes

The publicity surrounding the workshop raised the visibility of computer model evaluation within a variety of organizations (e.g., NCAR, LANL, and NRC) and within the general population of scientists, applied mathematicians and statisticians. (See "Researchers Look to Statistics in Quest to Quantify Uncertainty" by Barry Cipra in SIAM News January/February 2000 and "Revealing Uncertainties in Computer Models" by Cipra in Science 11 February 2000.) The interactions during the workshop appear to have already fostered a number of new cross-disciplinary collaborations (for example LANL statisticians began collaborating with LANL accelerator modelers to evaluate a complex model to predict accelerator performance, epidemiologists, statisticians, and applied mathematicians came together in a summer workshop to evaluate epidemiological models, and UC Berkeley statisticians and those at NCAR joined to evaluate

climate models). In addition, the positive reactions to the workshop have motivated CATS to consider several new initiatives to foster research in the area of model evaluation.

The presentations and discussions reinforced the observations that motivated the workshop. The diversity of the presentations emphasized that computer models are found in a variety of scientific fields (e.g., meteorology, oceanography, engineering, biology) and policy applications, and have many different formal structures (e.g., differential-equation-based, discrete, deterministic, stochastic). These models are often complex in the sense that they attempt to resolve a large number of relationships, many of which are highly non-linear, often with positive and negative feedback. Indeed, the first two presentations, one on storm systems (Section 3.1.1) and one on wildfires (Section 3.2.1) illustrated well the nature and consequences of model complexity even when the underlying theory is well developed. .

The workshop discussions underscored that current computer model evaluation practice is typically inadequate, sometimes grossly so. Possible explanations were identified as the enormous difficulty of the task, the inadequate resources, the need for new evaluation strategies and tools, and the incentive structure of the scientific community. Nevertheless, the workshop revealed cases where difficult obstacles were squarely addressed and at least partially overcome. For example, the transportation model (Section 3.3.1) was used change operations and the changes proved successful. The immunology model (Section 3.4.1) was used successfully to explain surprising real-world data.

Another issue actively discussed during the workshop was the need for a general framework for computer model evaluation. In the opening address to the workshop a taxonomy of models was proposed (Section 2). This taxonomy described seven kinds of models, organized in part by whether the models were deterministic or stochastic and by whether they were meant to represent the mechanisms by which empirical phenomena operated or meant to link inputs to outputs in a manner that maximized the fit. Many participants felt that an effective taxonomy was a requirement in developing a common language with which to speak about model evaluation.

During the course of the workshop, four themes emerged that provide a focus for model evaluation research.

1. *Context Is Critical* --- Model evaluation is not done in a vacuum; context is fundamental in defining what a "good" model is and how good a model needs to be. Context, in the sense of subject matter knowledge, is obviously essential in evaluation as well as in model construction. However, context has other dimensions for evaluation, such as whether the goal is scientific understanding, forecasting, training, and/or decision-making. Context will also drive the requirements for the amount and kind of precision necessary.
2. *Available Data Are Often Inadequate* --- The availability of useful "ground truth" data varies enormously. In some extreme cases (the wildfire control

model being one) the prospect of germane data seems dim, given the limitations of current technology and resources. Nevertheless, with reasonable resources and better designs, data that are far more useful could be collected for many models, as seen in the discussions of the meteorology, transportation, and immune system models.

3. *Uncertainty Analysis is Crucial* --- There is a vast, even bewildering, range of sources of uncertainty and variability in computer models. All of the sources of uncertainty listed in the *"What makes model evaluation difficult?"* question were manifest in one or another of the four workshop examples. Current representations of simulation uncertainty, even when provided are at best incomplete. Urgently needed are methods, algorithms, and software tools that can incorporate the multiple sources of uncertainty to produce the needed overall uncertainty calculations. Combined assessment of the various types of uncertainty is a formidable challenge in most situations, but clearly critical when model predictions are used in high-stakes decision-making.
4. *Better Data Reduction Methods Are Essential* --- The key strategies in model evaluation require making comparisons between model output and "ground truth" data and/or between different models. In both cases, however, existing methods that might facilitate such comparisons are too often overlooked (this point received considerable attention during the meteorology model discussion). There is a vital need for new methods that can be effectively applied to very large, high dimensional output/datasets, such as those from ocean models and satellite imaging data. The difficulty is, in part, finding data reduction procedures that do not obscure the scientific features of interest.

1.2 Research Roles and Programmatic Needs

Model evaluation is a cross-disciplinary exercise in which the interactions between modelers, applied mathematicians and statisticians are critical. Each of the examples in Section 3 highlighted that the underlying science, numerical methodological issues and data collection and analysis are inextricably interwoven. Statisticians and applied mathematicians can play a number of important roles in improving computer model evaluation: as the developers of new and better methods and tools, as the conduits by which statistical and mathematical technology are transferred to modelers, as consultants when the models are being developed and evaluated, and as full team members in a given modeling enterprise.

The evaluation of complex computer models is a rich source for new and demanding problems in statistics and applied mathematics. Statistics and applied mathematics, as disciplines, will directly benefit through the invention and utilization of new techniques for model evaluation emerging from research in Bayesian methods, uncertainty quantification in deterministic models, designing field experiments to match computer experiments (and vice versa), data collection

schemes, data reduction methods, data mining, and imaging and statistical visualization. In addition to the research, statisticians and applied mathematicians must meet the challenge of transferring these methodologies to the model developers and model users.

Fostering meaningful interdisciplinary collaborations has been of serious and widespread concern. Mechanisms to encourage young statisticians and other scientists to work in interdisciplinary environments need to be maintained and expanded. For example, it would be useful to have regular mechanisms by which young statisticians could be placed in modeling environments, such as the national laboratories. Similarly, young scientists could be placed in statistical environments such as major statistics departments. Another possibility would be to automatically build into funding for computer model development the resources to support serious model evaluation efforts (particularly those models used in decision and policy support). Currently, model evaluation is too often an afterthought, so that important opportunities are missed, and for those that remain, too little time and money are available. Given the enormous investment in developing the models and the importance of the decisions they effect, it is imperative that model evaluation be recognized as a critical component of the model development process, and that it be supported appropriately.

2. KEYNOTE ADDRESS BY DR. WILLIAM PRESS

Dr. Press provided workshop attendees with a lively and provocative keynote address. He stimulated considerable discussion by proposing the following taxonomy of computer model types:

1. "Accurate" models of deterministic physical phenomena with "accurate" input conditions
2. "Accurate" models of deterministic physical phenomena with "statistically accurate" input conditions
3. "Statistically accurate" models of non-deterministic physical phenomena
4. "Accurate" or "Statistically Accurate" models of emergent physical phenomena
5. "Phenomenologically Accurate" models that are not even statistically accurate
6. "Phenomenologically Interesting" models
7. "Video games" as models

Dr. Press noted that model evaluation requirements and issues vary depending on the model type. Accurate deterministic models (types 1&2) are models where the physics is well understood. He gave as examples static civil engineering

models of bridges and dams, weapons codes, and "well-behaved" hydrodynamics models. He noted that for these models a single model run could be compared to data, using an appropriate norm. Classic issues associated with these models include whether all the physics is captured in the model, the effects of truncation vs. round-off errors, and the choice of what numerical model output to compare to data. He noted that where the input quantities are inherently stochastic or poorly known (type 2) there is also an issue of how to quantify model uncertainty from this source.

Statistically Accurate Models (type 3) include those treating innately statistical phenomenon, classical chaos, and having seriously unknowable initial conditions. Examples include turbulent fluid phenomena and climate models. These models might be conceptually deterministic, but for ensembles, not individual realizations. However, it is generally computationally impractical to make many runs. Press suggested that a theory of sparse Monte Carlo simulations might be useful. He identified additional issues in this setting, including how to determine which runs to make and how to establish metrics for evaluation both for model-to-model and model-to-data comparisons. In particular, how might we formalize the typical "eyeball" comparisons between model and data?

Emergent Models (type 4) capture the desired macro-phenomenology by describing an underlying micro-phenomenology that results in the desired emergent phenomena. These models need not look like "real physics." Examples include statistical mechanics, smooth particle hydrodynamics (SPH), cellular automata hydrodynamic, and traffic flow modeling. Press warned that there is no meta-theory of emergence and using these models may leave one vulnerable to self-deception.

Phenomenological Models (types 5 and 6) capture qualitatively identifiable phenomena, e.g. turbulent intermittency, traffic jams and epidemics. These models can be "inaccurate" in almost any statistical sense, yet be extremely useful. They can be used to train human actors who can then quickly adjust to the actual phenomenology (e.g., fire fighting or war games). Issues for these models include how to map "fields of data" into "phenomena" or "events" and how to summarize the behavior (deterministic or statistical) of these phenomena.

Dr. Press concluded by suggesting a "space of models" and proposing a calculus for this space. He maintained that we seem to have an intuitive idea of such a space: codes can include *more* or *less* physics, models can be more *overlapping* or more *independent*, and we can envision a nested sequence of finer zoned codes. Dr. Press noted that intuitively we have the idea that a model result can be validated by *sampling* over the space of models and that when *different* codes agree, then they are both more likely to be accurate. He also noted that the more physics added, the more trustworthy the answer; ultimately, if we could compute "like nature" we would have the right answer. The claim of "the more physics, the more trustworthy" was actually disputed by the wildfire modeler and others in later sessions.

3. THE PRESENTATIONS

3.1 Mesoscale Modeling of Storm Events

3.1.1 The Scientific Problem and the Model
Presenter: Robert Fovell, University of California at Los Angeles

For the practice of atmospheric science, computer simulation models have become a key element in a wide variety of scientific and policy-related research. Applications can range in scale from particular chemical reactions to the climate of the entire earth. This presentation addressed mesoscale modeling of precipitation events in the Los Angeles Basin. The goal of the simulation was to provide model output, including but not limited to precipitation data, to be used as input to hydrological models for the investigation of streamflow and runoff in the region. As part of a computer model evaluation, it was decided to compare model output to data from one particularly strong precipitation event that occurred on 7-8 February 1993.

The computer model employed is known as MM5, or Mesoscale Model Version 5, a joint effort of the National Center for Atmospheric Research (NCAR) and Penn State University. MM5 is initialized using atmospheric observations and solves the equations governing physical, thermodynamic and microphysical processes within a three-dimensional domain that is subdivided into grid volumes. MM5 is a very complex modeling system. It includes many parameterizations that attempt to represent processes rendered unresolvable owing to temporal or spatial resolution limitations, such as the generation and dissipation of small turbulent eddies or the aggregation of cloud droplets into raindrops.

Precipitation fallout can be quite variable across the region, with the largest totals often recorded in the vicinity of mountain slopes. Therefore, accurate predictions of precipitation fallout depend upon an adequate representation of local orography, and this demands high spatial resolution. On the other hand, accurate handling of the frontal movement, including the timing of its arrival and speed of passage, is also of paramount importance, and this requires us to employ model domains sufficiently large to capture the front's parent storm system throughout the model simulation. This daunting combination is addressed with grid nesting. We used three nested domains having horizontal resolutions of 36, 12 and 4 km, respectively. The outermost grid crudely captures the entire storm system over the forecast period while the innermost domain concentrates the highest resolution in the local basin.

Two simulations were undertaken, both commencing at 4 PM local time on February 6th, approximately 24 hours prior to the onset of heavy rain. The lead-

time gave the model a chance to "lock on" to the storm and guide it towards the LA area. The first, or control, simulation was started with observations valid at 4 PM local time but made no use of observations collected subsequent to 4 PM. The second of the two simulations employed a technique known as Four Dimensional Data Assimilation (FDDA), a way of incorporating the subsequent observations into the model as it runs. The FDDA simulation represented an attempt to keep the model consistent with reality as it unfolded. Both simulations were integrated for 48 hours, fully spanning the time required for the storm to pass through Southern California.

Ideally, output from each simulation would have been compared to observed 48-hours precipitation accumulation over the basin collected at local range gauges. However, the gauges came in different types and styles, having different capacities and maintainers, and even incommensurable data collection intervals. Further, the gauges were relatively few in number, not optimally distributed, and often not sited in the areas of greatest interest.

Figure 1 shows the topography of the Los Angeles region, while Figure 2 depicts the 48-hour precipitation accumulation in the control run's innermost domain, superimposed (black contours) on the local topography. The largest totals are found on the mountain slopes that happened to face the wind during the storm passage. The front passed from west to east, but in the hours prior to its passage, a strong southerly flow entered the LA basin. This flow pushed copious amounts of moisture up and over the basin's mountains from the south, producing rainfall that accounted for a significant fraction of the 48-hour total.

The locations of some of the rain gauges in the LA area are also superposed on Figure 2. Gauges too close to the boundaries of this domain have been excluded. It is worth noting that there are forty times more surface grid points than gauges in the innermost domain. Figure 3 shows a scatterplot of model predicted versus recorded rainfall for the control run. The predictions were made by interpolating the model output to the rain gauge sites; a task made more daunting by the rapid variation of elevation in the vicinity of some of the gauges (especially those receiving the most precipitation, since elevation is clearly a major contributing factor). While the overall trend seemed acceptable, the control run was judged to have produced excessive precipitation overall, especially on the mountain slopes that received the largest totals. Figure 4 shows that the FDDA run, in contrast, was found to under-predict precipitation at most locations, especially the interior basin locales at which the rainfall totals were relatively smaller. A comparison of the two simulations suggested that the FDDA technique resulted in a slower moving front with weaker southerly flow ahead.

Comparing the two simulations was made very difficult by the problem of having such a wealth of information. There were many gridpoints, prognostic variables, and time steps within an integration. In short, the two simulations were found broadly to differ, but how could those discrepancies be properly unpacked and

presented? What, why and where did the two runs diverge? It was not a problem for simple summary statistics or histograms or statistical tests of these characterizations

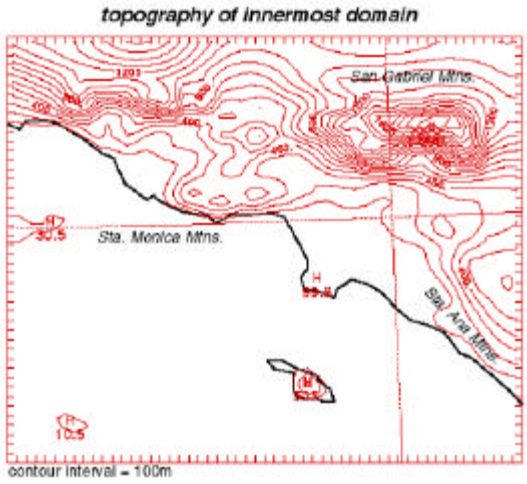


FIG. 1. *Topography of the MM5 model's innermost domain, centered on the Los Angeles basin. Contour interval is 100 m.*

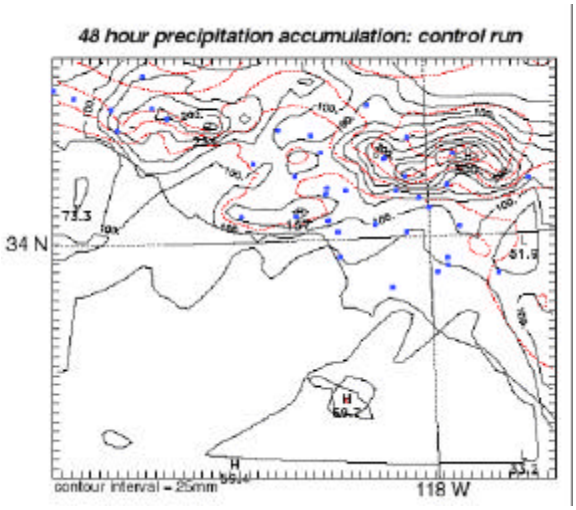


FIG. 2. *48 hour precipitation accumulation in the innermost domain for the control run. Contour interval is 25 mm. Dots mark the locations of rain gauges used in Fig. 3.*

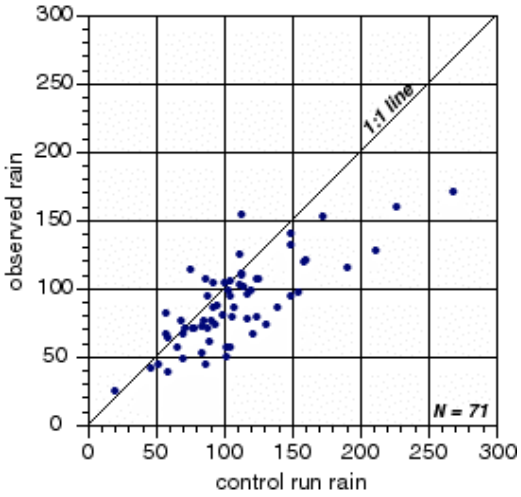


FIG. 3. *Scatterplot of observed versus predicted precipitation accumulation, using data from the control run. The 1:1 correspondence line is also shown.*

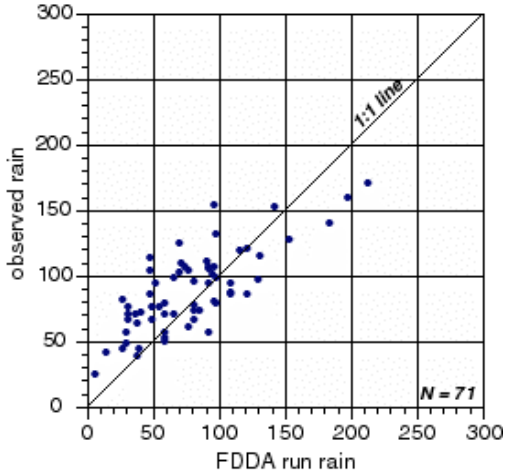


FIG. 4. *As in Fig. 3, but using the predictions from the FDDA simulation.*

In pointed contrast, the comparisons of model forecasts with reality were greatly hindered by the paucity of data. Precipitation, for example, is the end product of a huge number of interactions and processes, many occurring over time and space scales beyond the reach of our instruments. That made it more important to utilize what data there are, from whatever source, to find clues that may help identify the source(s) of forecast errors. This also was not a problem for simple summary statistics and conventional tests

3.1.2 The Statistical Assessment
Presenter: Richard Berk, University of California at Los Angeles

In the simple terms, the problem to be solved was, how does one determine which of these two simulations is "right?" The answer, if any, should depend upon the particular context in which the simulation results were going to be used.

In this instance, the goal of the modeling was to provide precipitation information to hydrologists for their research. Thus, where the rain fell was as important as how much. Among other things, the hydrologists were interested in potential flooding. Heavy rain falling on one side of the Los Angeles coastal mountains would send the water into a different watershed than rain falling on the other side of the Los Angeles Coastal Mountains. For the hydrologists, a certain level of modeling precision was essential. Errors of several centimeters in total rainfall for a given storm, for instance, could be devastating. It might mean the difference between a serious flood and rainfall that could be contained within the existing storm water infrastructure. Moreover, ad hoc "fixes" for the modeling would be unacceptable. Since the hope was to use the model for the wide range of storms that can reach the Southern California coast, an ad hoc fix forcing a simulation to "fit" for one storm would not quarantee that it would "fit" for others. So it was essential to understand the causes of any modeling errors and to make structural changes in the model as necessary.

Given this context, there were a number of difficulties to be faced in evaluating the models. The existing data were sparse and collected from rain gauges that were not ideally located. For example, there were very few rain gauges at higher elevations, precisely where rainfall is likely to be the heaviest and most heterogeneous. There were initially few hunches about how the model outputs differed or why. Complicating matters was the common observation that there could be lots of things wrong with the models. There was no idea about which things were likely to be most wrong.

The strategy that followed was to see what could be learned from a comparison between the output from the two models and also between both sets of output to what "ground truth" data existed. In the interest of time, only the simulation comparisons were addressed at the workshop. Here the idea was to link significant disparities between the output from the two simulations to variables that might suggest what was going wrong (e.g., elevation). In so doing, we proceeded from very simple statistical summaries that disgard lots of information to more complex statistical summaries that discard less.

This exploratory data analysis perspective, trying to understand how and why the two outputs differed, is more appropriate in this context than confirmatory statistical inference (tests or confidence intervals). Both simulations were deterministic, so differences in output were also deterministic. Certainly the data were stochastic, but the measured precipitation totals were both spatially and

temporally dependent and came from a convenience sample of rain gauges. In addition, the nature of the measurement error was unknown. Consequently, we saw no way properly to map conventional statistical inference onto the problem at hand. Ritual statistical inference would have enlightened no one, even though statistical tests on deterministic models are de rigeur in some climate research circles (e.g., chapters 7 and 8 in von Storch and Narvarra, 1999.)

We began with the simple descriptive statistics shown in Table 1. It was apparent that means differed enough to matter. In particular, the control run predicted more rain on the average, and the simple difference was large enough to matter in scientific and policy terms. Histograms of both sets of outputs revealed strong skewing to the right. Most of the predicted rainfall from the two simulations clustered at smaller rainfall amounts. But the FDDA run had many more observations piling up around zero.

A scatterplot constructed from the two sets of output showed a strong linear association and indeed, the correlation between them was 0.93. It would have been tempting to stop the analysis at this point, consistent with the kinds of model evaluations commonly found in the scientific literature. The conclusion would have been simple: the pattern of output from the two models is much the same, but the FDDA model predicts a bit less rainfall on the average.

However, the fact that FDDA histogram showed more predicted rainfall amounts clustering near zero suggested that disparities between the two sets of outputs might have had an important spatial component. Given the topography of the Los Angeles basin, longitude, latitude and elevation should be related to precipitation. A simple linear regression was applied in which the arithmetic differences between two model outputs were regressed on those three variables. The results are shown in Table 2. Overall, the results suggested that as one moves to the northeast and to higher altitudes, the simulations are more in agreement.

But the regression fit was poor and a wide range of model diagnostics indicated that the specification was a least very incomplete. For example, sliced inverse regression clearly indicated that we needed far more than a 1D structure and that in fact, 3D structure would perhaps be required. One implication was that we needed to fit a surface that allowed for non-linearities and product variables. We eventually arrived at the results shown in Table 3.

TABLE 1
Univariate Summary Statistics (Simulation output is in centimeters of rainfall.)

Variable	Sample Size	Mean	Standard Deviation	Min.	Median	Max.
Control Run	2430	8.3693	4.3328	2.33	7.105	30.04
FDDA Run	2430	5.0694	4.952	0.36	2.83	27.58
Control - FDDA	2430	3.2999	1.8505	-6.67	3.58	7.76

TABLE 2
Linear regression of the simulation differences on location
(N=2430, R^2 =0.23)

Variable	Coefficient	Stand. Error	t-value
Constant	5.39	6.15	0.88
Longitude	–0.20	0.05	–3.84
Latitude	–0.76	0.11	–7.04
Elevation	–0.0012	0.00011	–10.32

TABLE 3
Linear Regression allowing for 3-Dimensional Structure
(N=2430, R^2 =0.61)

Variable	Coefficient	Stand. Error	t-value
Constant	–12137.5	966.149	–121.563
Longitude	–167.15	16.16	–10.34
longitude sq	-0.76	0.07	–10.68
Latitude	127.62	18.36	6.95
Latitude sq	-2.78	0.27	–10.26
Elevation	12.96	2.11	6.13
Elevation sq	0.00	0.00	-0.74
Lat X Lat	–0.50	0.15	–3.27
Lat X Elevation	–0.36	0.06	–5.95
Long X Elevation	0.10	0.02	6.14
Long X Lat X Elev	-0.00	0.00	-5.95

An examination of the final response surface provided a number of insights. For example, the simulation output seemed to differ most as the storm hit the front edge of the coastalrange. More generally it now seemed clear that the disparities between the models had important scientific information, that the disparities had strong spatial relationships, and that these might be explained by differences between the two simulations in the angle at which the storm approached the coast.

What might be concluded from this exercise more generally about the role of statistics in computer model evaluations? Perhaps most important, there was a need to approach the virtual world produced by the models with the same care that one would employ for the real world. This implies the need for careful description as the foundation for understanding. In the eyes of many statisticians this is "blue-collar" work in which formal models to capture uncertainty are sorely missed. However, if formal models are applied without sufficient understanding of the virtual world being constructed, those models will be irrelevant or worse. Something that is not worth doing, is not worth doing well.

3.2 Wildfire Modeling

3.2.1 The Scientific Problem and the Model
Presenter: Rodman Linn, Los Alamos National Laboratory

In order to facilitate better decision-making regarding wildfire behavior, response, and effect, people have been working to model wildfires for more than a half century. The majority of the operational wildfire models currently in use are empirical and have been developed based on a limited number of idealized experiments. Because of the limited number of wildfire regimes used to develop these empirical models, they are not appropriate in many wildfire circumstances. These empirical models are also limited in their ability to model the strong coupling between the many physical processes that exist in a wildfire (heat transfer, chemical reactions, moisture extinction, buoyancy induced turbulent flows, flow through canopy, etc.), and are thus limited in their ability to predict emissions and ecological effects of wildfires.

Los Alamos National Laboratory (LANL) is developing a wildfire model, FIRETEC (1), which is based on simulating the physical processes that control wildfire behavior. FIRETEC is designed to capture the combined effects of small micro-events on the macro-scale fire behavior. FIRETEC can be used to model fires and their effects in most realistic wildfire circumstances (complex terrain, variable wind conditions, nonhomogeneous vegetation), including those that are impossible to simulate accurately with current operational models. This is because FIRETEC is based on conservation of mass, species, momentum, and energy and thus captures the driving physical processes of wildfires.

The framework of FIRETEC is a set of coupled transport equations. These transport equations (written in the form of partial differential equations) incorporate time and space history into expressions for momentum, internal energy, gaseous species concentrations, turbulence kinetic energy, and fuel moisture depletion. An example of one of these transport equations is given in symbolic form in Equation 1. Equation 1 describes the transport equation for internal energy for the combined gas phase in the presence of a wildfire.

$$\begin{aligned}\frac{\partial(Internal\ \ Energy)_{gas}}{\partial t} = \ & (mean\ \ flow\ advection) \\ & + (Diffusion\ due\ to\ turbulence) \\ & + (Net\ radiation\ source\ to\ gas) \\ & + (Net\ convective\ heat\ exchange\ to\ gas) \\ & + (\ Internal\ energy\ source \\ & \qquad due\ to\ chemical\ reactions)\end{aligned} \tag{1}$$

The various terms depicted on the right side of equation 1 are strongly coupled with terms in other transport equations as shown in the following table.

Term in Internal Energy Equation	Other transport equations that this term is directly coupled to
Mean flow advection	Equations for Momentum
Diffusion due to turbulence	Turbulence kinetic energy equation
Net radiation source to gas	Gaseous species concentration equations
	Internal energy of the solid fuels
Net convective heat exchange to gas	Internal energy of the solid fuels
	Equations for Momentum 3 directions
Internal energy source due to chemical reactions	Turbulence kinetic energy equations
	Gaseous species concentration equations
	Internal energy of the solid fuels

This coupling makes it very difficult to isolate and evaluate specific processes that are occurring in a wildfire. The coupling also makes it very easy for a small error in models of a single process or quantity to propagate to the representations of other processes or quantities. Inaccuracies in input data are also prone to have very complex ramifications in such a complex model, and the outcome of such a model will often be altered in unforeseen ways. Therefore, a systematic statistical uncertainty analysis is critical for understanding the uncertainties of this type of physics-based model, and for understanding how sensitive the results are to inaccuracies in input data.

Physics-based wildfire models have the ability to predict very detailed behaviors such as individual wind gusts, specific details of fire-line shape, and locations where vegetation is not completely burned. The presence and nature of some of these fire-behavior details helps give the models like FIRETEC credibility because qualitatively similar features occur in nature. However, the specific location where there is a patch of fuel which is not burned, the specific point on a fire-line that sticks out ahead of its neighbors, or the moment that a gust erupts from a fire are all details that are very dependent on unresolvable input details (the specific vegetation configuration, the small in coming wind gusts, etc.) These unresolvable input details are not measured accurately but they will have large effects on precise nature, location, and timing of some on the fire-behavior details.

To further complicate matters, it is very difficult to develop a wildfire model without adequate real wildfire data for comparisons and validation. In order to get data that are useful for validating and developing wildfire models, data acquisition methods must be employed under realistic wildfire conditions and must monitor a wide variety of physical quantities (temperature, velocities, etc..) The difficulty in obtaining these data is partially in developing instrumentation that will capture the data at the proper detail but also in finding "realistic conditions" under which the data can be collected. It is very difficult to get adequate comprehensive

instrumentation deployed on a true wildfire due to the unpredictability of a wildfire. A initial step at acquiring "wildfire data" can be made by taking data from controlled burns, but even under some of the best planned experimental controlled burns, there are a limited number of conditions under which the experiment may be performed. It is possible to perform experiments that help to validate particular aspects of wildfire models, but it is difficult to confidently couple the parts of the model without having a way to validate the overall wildfire behavior under a variety of conditions.

3.2.2 The Statistical Assessment

Presenter: Frederic Schoenberg, University of California at Los Angeles

There are perhaps two broad types of model evaluation: internal and external. Internal model assessment is the evaluation of the structural relationships prescribed by the model. For instance, a model for the spread of wildfires may be based on relationships between spread rate and variables such as wind, temperature, vegetation, fuel moisture, precipitation, and so on. Some of these relationships may be examined individually. Such internal model evaluation is most often done using laboratory experiments, where each of the variables can be carefully controlled and measured. Unfortunately, the relevance of such evaluations can be questionable because of scaling problems. For example, most experiments on fire behavior involve fires of sizes on the order of inches; extrapolation of such results to multi-acre forest fires may be spurious.

Alternatively, one may examine the external features of the model: how well the model describes the broad features of the phenomenon in question. For instance, one may observe actual forest fires and see if their behavior in certain aspects of interest – e.g. temperature, burn pattern, flame angle, and spread rate – agree with the model. Unfortunately, data for such evaluation is typically quite sparse.

The above relates to comparing models to observations. Also relevant to model assessment is the comparison of models to models. Different models may be simulated repeatedly, and their results compared to each other and/or to observations.

When comparing two objects (such as outputs from two models, or model output and observations), several statistical tools may be helpful. Certainly the importance of basic data-analytic techniques such as regression should not be overlooked. Even in a deterministic setting, it may be useful to fit a line or other curve to the data as a way of summarizing output. Stochastic model assessment techniques, such as simulation methods and likelihood methods, have proven extremely important in the evaluation of stochastic models. However their usefulness is debatable for assessing deterministic models, for which likelihoods are undefined and simulation outputs under the same initial conditions are identical.

A problem that occasionally arises in the evaluation of deterministic computer models is that of finding a simpler model which closely approximates the computer model. Such an approximation, which may be called a Statistically Equivalent Model (SEM), may be useful for various purposes including the description, inversion, and simulation of the model. The coefficients of a SEM are often easily interpretable and thus may highlight important features of the computer model. Further, while complex computer models are often costly to simulate and difficult if not impossible to invert, such may not be the case for a SEM.

One specific technique which may be useful is to approximate a complex computer model by a locally linear, space/time-invariant filter (Schoenberg et al., 2000.) By locally linear, we mean that the output perturbation resulting from a perturbed input is a linear function of the input perturbation. By space/time invariant, we mean that the relationship between perturbed inputs and outputs does not vary with location and time. The idea is that some complex computer models, though perhaps highly nonlinear, may be locally nearly linear. The benefit of approximating such models in this way is that local linear systems are very easy to simulate, invert, and interpret.

At present, rigorous evaluation of complex, deterministic computer models is rarely done. Why not? The most common reasons cited are limited data, scaling problems, and computational burdens. However, the main problem may instead be cultural. Many prominent statisticians traditionally have a cautious view of models applied to real-world phenomena while many of the most visible physical scientists readily embrace them. Moreover, the rewards in the two fields may reinforce the model skepticism of statisticians and the model dependence of physical scientists. Ways need to be found to help bridge this cultural divide. Without better communication and collaboration, improved formal techniques are beside the point.

3.3 Transportation Modeling – Design and Evaluation of Traffic Signal Timing Plans

3.3.1 The Scientific Problem and Model
Presenter: Nagui Rouphail, North Carolina State University

The development of efficient signal timing plans for urban traffic networks is a continuing challenge to traffic analysts and engineers. Flows on these networks, even small sub-networks, are highly complex: they encompass a variety of vehicles (autos, trucks, buses), pedestrian-vehicle interactions, driver behavior, and an assortment of network conditions (lane arrangements, stop signs, parking lots, one-way streets). Moreover, the traffic demands on the network are highly variable (minute-to-minute, hour-to-hour, day-to-day, month-to-month) as are many of the movements (even legal ones) of vehicles and pedestrians.

Over time and through experience and modification traffic managers have developed signal control strategies to respond to these conditions. In recent years

they have been assisted by traffic models, sometimes oversimplified, that can generate signal control strategies (see Click and Rouphail, 1999, for a review of a number of these).

At the same time, there has been a steady development of microsimulation computer models that simulate traffic under a complexity of conditions, including traffic signal settings. One such, Corridor Simulation (CORSIM) has been adopted by the Federal Highway Administration (FHWA) as the quasi-official platform upon which to gauge traffic behavior and compare competing strategies for signal control before implementing in the field.

This leads to two crucial questions:

1. How well does CORSIM reproduce field conditions?
2. Can CORSIM be trusted to represent reality under new, untried conditions (e.g., revised signal timing plans)?

To address these questions we undertook a case study, with the cooperation of the Chicago Department of Transportation (CDOT), and the Urban Transportation Center (UTC) of the University of Illinois at Chicago. The test-bed for the study is the network depicted in Figure 5. The case study involved data collection for inputs required to run CORSIM and to optimize, over signal plans, performance measures defined on CORSIM output, as well as for data to evaluate CORSIM's capability to model field conditions.

CORSIM is a stochastic simulator that moves vehicles second by second through a network. Three types of inputs are required:

1. Fixed and non-controllable inputs: The network in Figure 5 represents a set of urban streets within the City of Chicago. In CORSIM, streets and intersections are modeled as directed links and nodes, respectively. Specification of the network includes a set of fixed inputs describing the geometry (e.g., distance between intersections, number of traffic lanes, length of turn pockets), the placement of stop signs, bus stops and routes, and parking conditions.

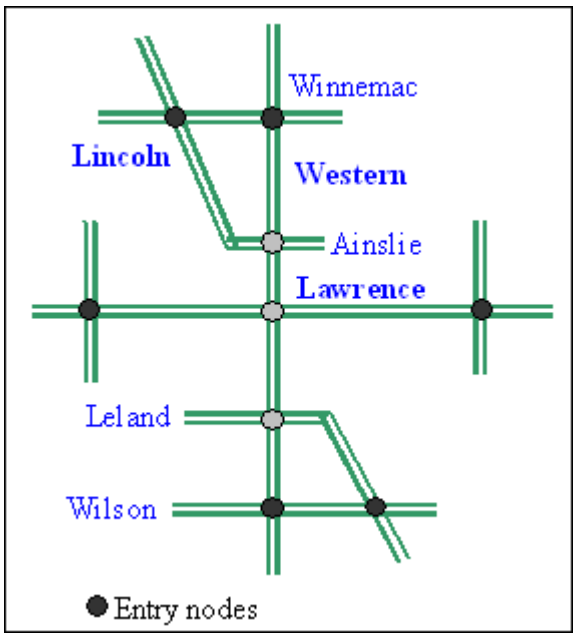


FIG 5. TEST-BED NETWORK.

2. Random and non-controllable inputs: Vehicles – autos, trucks, and buses – are generated by sampling interarrival time distributions at each entry node of the network. The interarrival time distributions are assumed to be independent (vehicle-to-vehicle, node-to-node), and may be different for each entry node. The designation of vehicle type – auto or truck – is made through independent Bernoulli trials with a fixed probability estimated from field data. Buses are treated according to their schedule and routes, with random dwell times at bus stops and random interarrival times at entry nodes.

   The behavior of the traffic is affected by additional random factors such as turn probabilities, driver characteristics (car-following behavior, lane changing maneuvers). Default distributions are provided in the CORSIM software for some of these (for example, driver aggressiveness) while others (such as turn probabilities) need to be estimated or specified..

3. Controllable inputs: For a signal study, the signal settings must also be specified:
   - cycle length – we assume a common cycle length for all signals
   - green times at each intersection – how long the signal is green for straight through movement, protected left turns,

- offsets – the difference in time between the start of the "green" through-movement at a signal and the time of the start of the green through-movement at a reference signal.

For the network of Figure 5 there are 22 signal parameters: 1 cycle, 13 green times, 8 offsets.

One hour of simulation, about 7500 vehicles total, takes about 1 minute on a PC with a Pentium II-500 Mhz. The status of each vehicle is updated every second. While each run is fairly quick, the need for many runs to deal with the substantial variability induced by the stochastic assumptions and the optimization of the signal parameters makes the time for an experiment non-trivial.

CORSIM comes equipped with an animation package that enables visualization of the traffic movements, a capability of great value in exploring the characteristics of the model and detecting problems and flaws. Besides the visual output, CORSIM provides aggregated (over selected time intervals such as the signal cycle) numerical output for each link: number of trips on each link, average link travel time, link queue time (the sum over vehicles of the time, in minutes, during which the vehicle is stationary, or nearly so), maximum queue length on each lane in the link, link delays (simulated travel time minus free-flow travel time, summed over all vehicles traversing the link). It is from these outputs that performance measures are taken.

3.3.2 The Statistical Assessment
Presenters: Jerome Sacks and Byungkyu Park, National Institute of Statistical Sciences

The data available for this case study were collected during mid-week morning and evening peak periods, one hour each. A platoon of human "counters" was employed to count vehicle arrivals (cars, trucks) at each boundary (entry) node for the entire one-hour period in the morning and evening. Turning movements at all links were counted: some links for short periods (15 minutes), some for one hour. Total vehicle flows and maximum queue lengths were counted on key internal links over the one-hour periods.

Input parameters were estimated from data as follows. The vehicle mix was estimated by observed proportion, as were the turning probabilities. The parameter of the interarrival time distribution at an entry node was estimated by a simple moment estimator of a parameter of a gamma distribution.

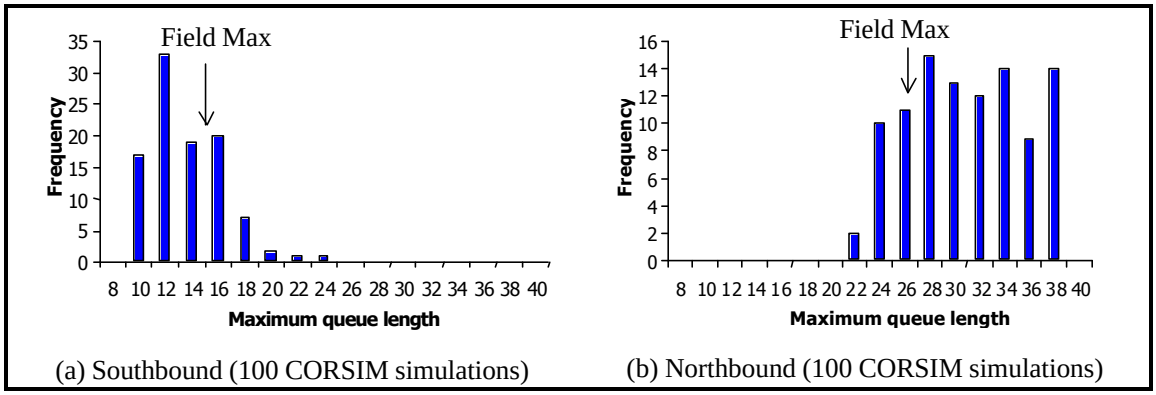
Using the estimated input parameters and the existing (base) signal plan we made 100 independent runs of CORSIM and viewed histograms of maximum queue length (MQL) for 6 key links (2 are plotted in Figure 6). This was done for the morning and evening periods. The observed field values are clearly well within the simulated ranges (as they were for the key links not shown in Figure 6) and tempts us to accept CORSIM traffic as a good representation of reality. But this would be

highly heuristic and subjective. Traffic is inherently variable, but the variability reflected in the histograms in Figure 6 may be excessive. The dependence between MQL and the data used to estimate the inputs to the simulator is not accounted for, nor are the dependencies among the MQL histograms for the various links.

Explicit in the above process is the selection of an evaluation function, the MQL. There are many potential candidate functions that could be considered, and the selection is somewhat arbitrary, but motivated by two considerations: (1) the feasibility of collecting the corresponding field data and (2) the connection between large MQL and potential for spillback and gridlock (nightmare conditions for traffic managers.

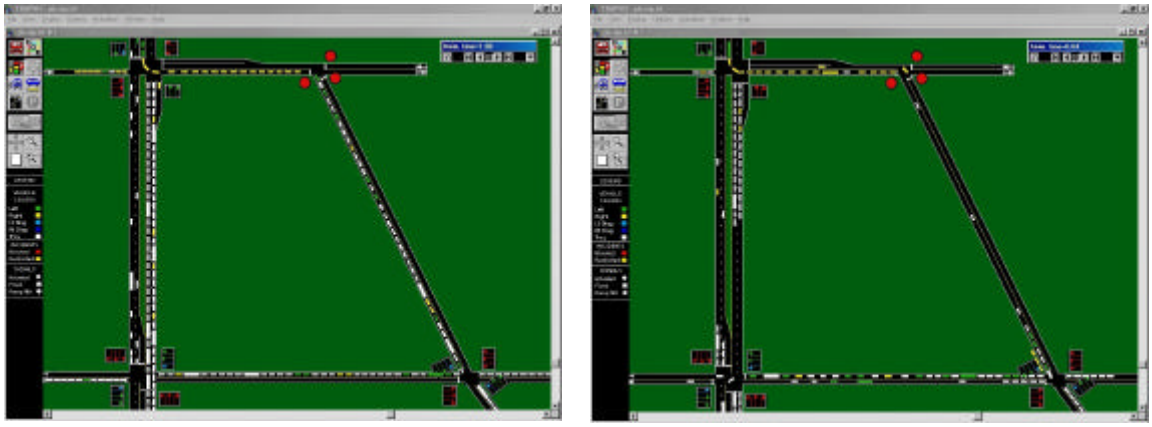
The presence of spillback or gridlock in several of the repeated simulation runs can be a sign of problems with the simulator. That presence may be numerically indicated by low throughput (number of completed trips) and by large "run-to-run" variance. For example, an optimal signal strategy (call it SS1) produced, in 100 runs, a mean queue time of 224.3 hours, a median of 180.9, and a standard deviation of 90.8 hours.

This high variability led to a close examination of the animation to uncover the circumstances leading to such large queue times. This in turn led to the snapshots (the 17:36:25 animation snapshot and the 17:23:30 one) in Figure 7. The cause was found to lie in the presence of a stop sign (at the upper right part of Figure 7) with overly long stop times. These are not found in the field because of the common practice of "rolling stops", at least when police cars are absent. Indeed, when the stop sign was altered to reflect this reality the results were striking: the optimum plan under the new circumstances had mean queue time of 122.5, median 111.2, a standard deviation of 34.4, and an absence of spillback.



(a) Southbound (100 CORSIM simulations) (b) Northbound (100 CORSIM simulations)

FIG 6. *MQL distribution from CORSIM and field at intersection of Western and Lawrence (see Fig. 5).*



(a) Snapshot taken at 17:36:25 (b) Snapshot taken at 17:23:30

FIG. 7. ANIMATION SNAPSHOTS FROM AN OPTIMAL SIGNAL STRATEGY (SS1)

To find optimum signal plans we adopted an objective function that minimizes a modified network queue time (MNQT):

$$MNQT = \sum_{i=1}^{L}\left[QT(i) \times \left\{ \max\left(1, \frac{MQL(i)}{SC(i)}\right)\right\}^{2}\right]$$

where: $QT(i)$ = queue time on link i, i= 1,2,…., L
$MQL(i)$ = maximum queue length on link i
$SC(i)$ = through signal capacity on link i, a function of the cycle and green time on link i.

Thus, we seek to minimize the network-wide queue time while penalizing overly long queues. The penalty avoids solutions that reduce queue time on a busy link at the expense of long queues on a less busy link. Optimizing this function over the 22 signal parameters was done via a genetic algorithm (Goldberg, 1989) applied to this stochastic optimization problem where the objective function is observed with error – the observation being a result of a single CORSIM run. An earlier effort via response surface fitting could not cope with the high degree of dependence among the offset parameters, the substantial run-to-run variability in the simulator, nor the non-smooth dependence of the objective function on the offsets. Having found an optimum this way, an additional 100 runs were made to compare the distribution of queue time of the optimum plan with other plans, especially the (current) base plan. These plots are shown in Figure 8.

To treat the uncertainties generated by the input data and the dual-use of the collected data for model inputs as well as for evaluation, we will need a more elaborate framework. Abayesian approach (for example see Bayarri and Berger 1999) could be a powerful one though potentially involving formidable calculations.

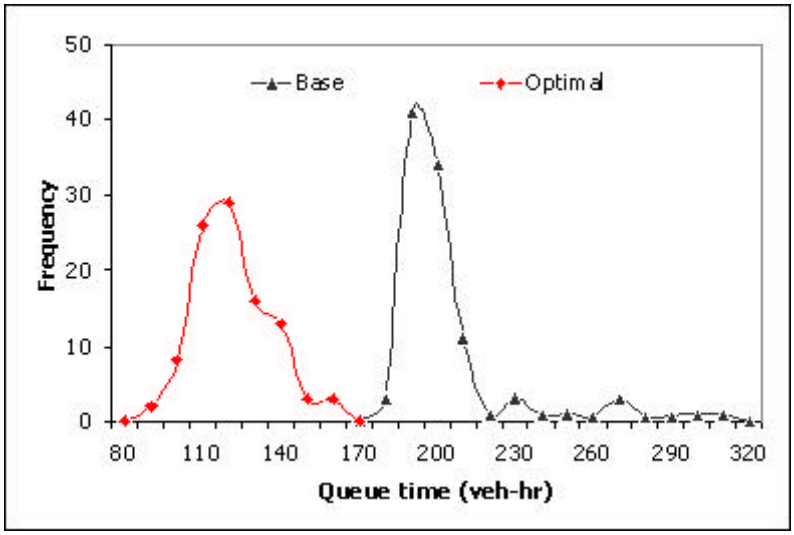


FIG. 8. *Queue time comparison between optimal and base case*

- designing and collecting field data is critical, difficult, and expensive,
- the simulator might be assessed under current field conditions, but changing the conditions –- for example, changing to an optimal signal plan –- will require a further field experiment to instill confidence that the simulator continues to reflect reality after the change. It would be poor advice to tell the Traffic Commissioner to change signal plans
- without an accompanying caution that a study following implementation is necessary. Robustness in "simulator" world is not necessarily robustness in the real world.

The use of visualization (in CORSIM, the animation) is important because it can quickly provide insight into difficulties and also assist in uncovering sources of trouble.

3.4 Influenza Modeling – Annual Influenza Vaccination

3.4.1 The Scientific Problem and the Model
Presenter: Alan Perelson, Los Alamos National Laboratory

In a typical flu season between twenty and forty thousand people die from the complications of influenza infection. Vaccination is our major weapon in protecting against flu, and each year tens of thousands of people get vaccinated. In the 1970's and 1980's two large clinical studies addressed the question of whether people vaccinated one or more times in the past were better protected than people getting a flu shot for the first time (1-3). Surprisingly, they reported conflicting

results: in some years it appeared that first-time vaccinees had better protection than repeat vaccinees, while in other years the reverse was the case. Smith et al. (4) have proposed and tested the antigenic distance hypothesis to explain the heterogeneity of repeated annual influenza vaccination. The antigenic distance hypothesis is illustrated in Figure 9.

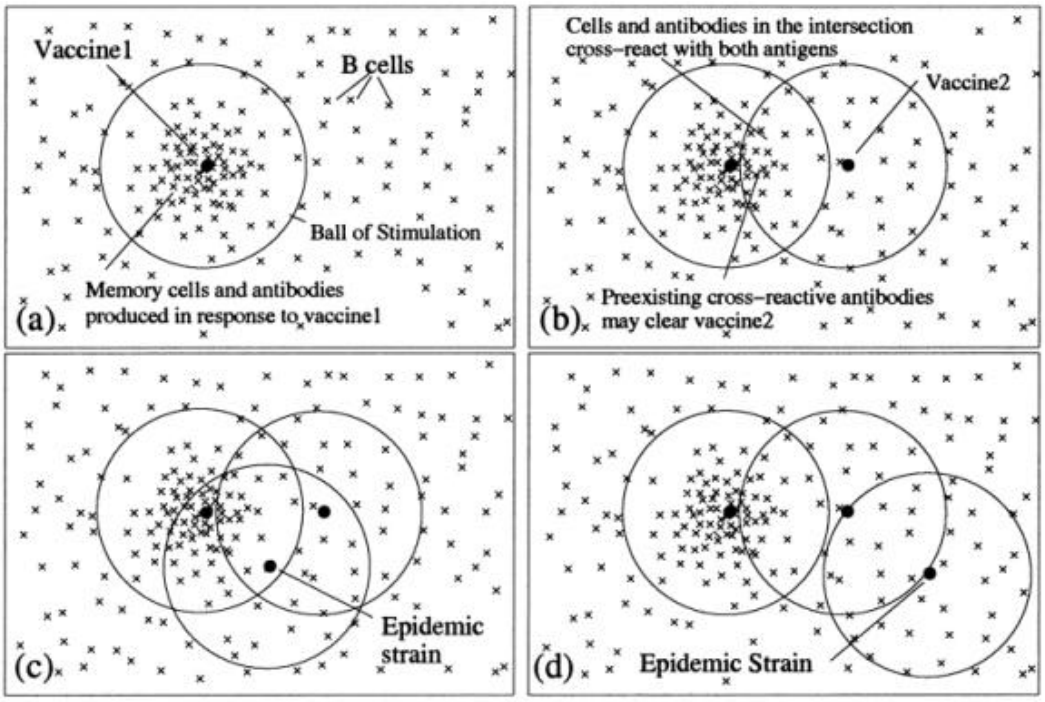


FIG. 9. *An illustration of the antigenic distance hypothesis. The affinity between a B cell or antibody (×) and an antigen (solid circles) is represented by the distance between them. Similarly, the distance between antigens is a measure of how similar they are antigenically. (a) B cells with sufficient affinity to be stimulated by an antigen lie within a ball of stimulation centered on the antigen. Thus, a first vaccine (Vaccine1) creates a population of memory B cells and antibodies within its ball of stimulation. (b) Cross-reactive antigens have intersecting balls of stimulation, and antibodies and B cells in the intersection of their balls. Those with affinity for both antigens are the cross-reactive antibodies and B cells. The antigen in a second vaccine (Vaccine2) will be partially eliminated by preexisting cross-reactive antibodies (depending on the amount of antibody in the intersection), and thus the immune response to Vaccine2 will be reduced (8, 9). (c) If a subsequent epidemic strain is close to Vaccine1, it will be cleared by preexisting antibodies. (d) However, if there is no intersection between Vaccine1 and the epidemic strain, there will be few preexisting cross-reactive antibodies to clear the epidemic strain quickly, despite two vaccinations. Note that in the absence of Vaccine1, Vaccine2 would have produced a memory population and antibodies that would have been protective against both the epidemic strains in c and d.*

The computer model simulates the antibody response to influenza vaccination and to exposure to an epidemic strain of flu virus. The model is agent based and considers a population of 10^7 B lymphocytes, the cells that secrete antibodies. In the model each B cell is characterized by having a randomly made receptor specified by a string of 20 characters on a 4 letter alphabet. Distances represented in two dimensions in Figure 9 are actually Hamming distances in 20-dimensional space in the model. When a B cell is stimulated it secretes its receptor as a soluble molecule called antibody. Antibodies that are partially matched to a flu virus (the antigen) bind the flu virus and lead to its elimination. In the model, flu viruses are also characterized by a similar length string and Hamming distance is used to compute the "antigenic distance" between a virus and antibody or B cell. if a flu virus partially matches the receptor on a B cell, which is represented in the model by a distance of 0 to 7 between the receptor and the antigen, we assume it can stimulate the B cell into reproduction and further antibody production.

The computer experiment considered two influenza seasons, one year apart, with four categories of individuals: (*i*) those never vaccinated, (*ii*) those who received "Vaccine1" (v1) at the start of the first influenza season and were not vaccinated for the second season, (*iii*) those not vaccinated for the first season but who received "Vaccine2" (v2) at the start of the second season ("first-time vaccinees"), and (*iv*) those who received v1 at the start of the first season and v2 at the start of the second ("repeat vaccinees"). All simulated individuals were challenged with epidemic virus two months into the second influenza season. The same v2 and epidemic strains were used for all simulated individuals, and v1 was varied. The antigenic distance between v2 and the epidemic (v2-e distance) was fixed at 2. Since cross-reactive distances vary between 0 and 7, this distance is "close," but it is not a perfect match. v1-e and v1-v2 distances varied between 0 and 7. The vaccine strains were nonreplicating, whereas the epidemic strain was able to reproduce. During each simulation if the viral load exceeded a "disease threshold" the simulated individual was considered symptomatic. Each experimental group contained 200 simulated individuals, and the attack rate within a group was defined as the proportion of the group in which the viral load exceeded the disease threshold. The results of the experiments, reproduced from (5), are shown in Table 4.

Table 4
Summary of experimental attack rates

v1-e distance	v1 only	v1-v2 distance for repeat vacinees							
		0	1	2	3	4	5	6	7
0	0.01**			0.00**					
1	0.46		0.06**	0.01**	0.04**				
2	0.87#	0.78#	0.37**	0.20**	0.19**	0.18**			
3	0.96#		0.74#	0.44*	0.36**	0.35**	0.38**		
4	0.99#			0.71#	0.50	0.45*	0.41**	0.50	
5	1.00#				0.66¶	0.46	0.47	0.50	0.50
6	1.00#					0.65¶	0.54	0.45*	0.54
7	1.00#						0.55	0.58	0.52

The fraction of individuals who develop flu symptoms, i.e., the attack rate, in the unvaccinated control was 1.0 (not shown). The attack rate for first-time vaccinees (v2-only) was 0.55 (not shown). Attack rates for repeat vaccinees and the v1-only groups are shown in the table. Groups marked with a ¶ or # had higher ($P < 0.05$ or $P < 0.01$, respectively) and groups marked with an * or ** had lower ($P < 0.05$ or $P < 0.01$, respectively) attack rates than did first-time vaccinees. Attack rates as high as 1.0 are due the large-dose experimental challenge of each simulated individual.

One immediate conclusion from the table is that repeat vaccination is always beneficial when given to previous vaccinees. This is illustrated that the attack rates in each row are lower for the repeat vaccinees than the individuals who received vaccine 1 alone. We also compared the simulation results to the two clinical trials that have evaluated the benefits of repeated annual vaccination, using data supplied by the Centers for Disease Control and Prevention, Atlanta, to estimate the antigenic distances between the various vaccines and epidemic strains in the study years. Figure 10 shows that the computer simulations had surprising agreement with the experimental observations.

Having a complex computer model that appears to provide insight into biological phenomena, how do we validate it? The immune system is a very complex system of which we have incomplete information and incomplete understanding. Thus, independently validating every underlying assumption in the model is impractical. Also, the computer code underlying the simulations is complex and needs verification. Once way to gain confidence in the computer code is to have someone write a completely independent simulation package based on the same biological assumptions and see if the same conclusions are

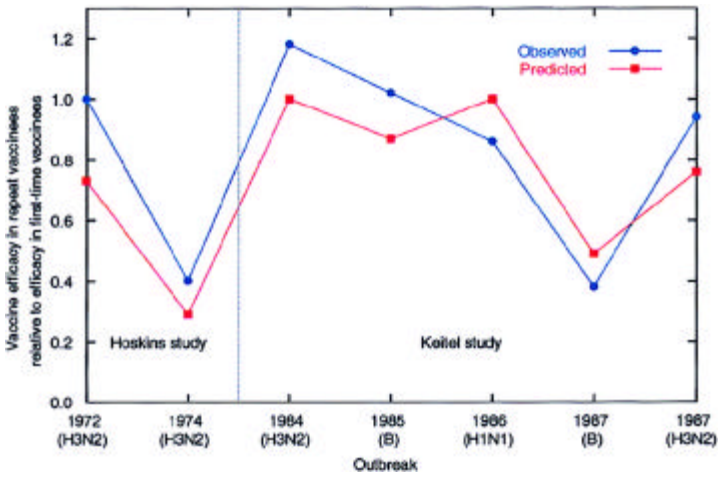


FIG. 10. *Observed vaccine efficacy in repeat vaccinees relative to the efficacy in first-time vaccinees, and predicted vaccine efficacy based on the antigenic distance hypothesis.*

reached. This also has the additional benefit of uncovering hidden assumptions made by the programmers when developing the code that might not have been adequately documented. Lastly, the results presented relied on a set of default parameters to describe the underlying biological processes. Parameter sensitivity studies are needed to validate the robustness of the results and to quantify the consequences of parameter uncertainty. This aspect of the validation process could be helped by finding means of speeding up the simulation or through the development of alternative yet simpler models that capture the main aspects of the antigenic distance hypothesis.

3.4.2 The Statistical Assessment
Presenter: Peter Bickel, University of California Berkeley

Immune system modeling raises in a crisp way most of the issues addressed in this workshop. The goals are on the one hand purely scientific, understanding how the second most complex system of the body works, and on the other hand very practical using models to predict the effect of vaccines, the progression of diseases such as HIV, etc. The antigenic distance model is particularly inviting to statisticians since it is purely stochastic. All objects, various types of B cells, antigens, antibodies are modeled as bit strings, their interactions and life histories are purely stochastic.

Verifying whether such models represent the detailed working of the immune system is difficult because the detailed data come from small animal experiments. This may change as other richer sources of information such as gene array data come on line. On the other hand seeing how well such a schematic computer model fits and explains apparent anomalies in epidemiological data is, as shown above, very feasible, and the measures of performance are clear. For instance, given such

a model, can one design vaccines with behavior predictable for given flu epidemics, at least *ex post facto*, knowing the type of the antigen causing this year's epidemic as opposed to previous years'?

Given predetermined intrinsic parameters, each run of the computer model produces attack rate data for set values of the three external parameters. What one is interested in is the behavior of the average attack rate over runs produced by the computer model as a function of vaccine-vaccine and vaccine-antigen distances, as well as the differences between model and observed (in epidemics) attack rates. All quantities observed have stochastic variability resulting from random seed differences in the model and in comparisons with real data unknown factors in the observations. Since one wants to make statements about all these parameters simultaneous inference is required. For studying the variability of the average attack rates over runs for fixed distance settings presumably the Gaussian approximation can be used. For studying the variability of average attack rates for different distance values variants of the Tukey or Scheffe methods (Miller, 1966) based on parametric bootstrap ideas (Efron and Tibshirani, 1993) are appropriate.

At the moment, computational pressures in this area seem minor compared to those, say in climate modeling. The antigenic distance model has 10-20 intrinsic parameters corresponding to affinity of antibody for antigen, rate of exposure of B cells to antigen, rate at which antigen stimulated cells divide, lifetimes of memory and plasma cells, and other rules of the model. In addition, there are three external parameters corresponding to antigen/vaccine and vaccine/vaccine distances. With intrinsic parameters predetermined by experience with animal experiments and three external parameters to vary it has been possible to run reasonable numbers of replicates and make confidence statements about average model outputs. However, computation time could become significant if one wants to make sensitivity studies involving the intrinsic parameters, or search for optimum vaccine strategies as a function of vaccine-vaccine and antigen-vaccine distances with more than two consecutive epidemic years.

For finding optimum vaccine regimes as a function of previously observed antigens and vaccine-antigen interactions, we need to maximize functions of combinations of distance settings, which are estimated from the averages over runs of outputs of the computer model. Such optimizations raise issues of sequential experimental design. For what combinations of distance settings should runs initially be made? How does one move to optimally search for the maximum? How many runs should be made at values of the parameter for which we count to estimate the function we are optimizing? This can depend on the model parameter values, etc. Methods for doing this have been considered in the statistical and engineering literature: see Box and Draper [1987], Kiefer and Wolfowitz [1952], and Kushner and Lin [1997] for surveys.

Stochastic cellular automaton models of the type considered above can be approximated by simpler stochastic models that might run more quickly. For

example, the antigenic distance model can be thought of as a continuous time, finite state Markov process, each of whose states is an array labeled by the 20 possible bit strings and the 400 pairs of bit strings. Corresponding to each bit string is a vector of counts recording how many unbound antibodies, antigens, memory and plasma B cells etc., corresponding to that bit string are present at that time. Corresponding to each pair of bit strings is the number of bound antigen-antibody pairs. Events such as binding, birth and death of cells occur with exponential lifetimes determined by the current state of the system and then change the current state according to a transfer matrix e.g., binding of an antibody with an antigen reduces the corresponding count entries at their two bit strings by 1 and adds to the count of bound antigen-antibody pair at the corresponding pair of bit strings. Of course implementation of this model is different; one would not wish to keep track of what happens at all bit strings or pairs of bit strings since most count vectors are 0's. In fact, given the definitions of affinities, one only needs to keep track of what is happening at various vaccine-antibody, vaccine-antigen distances. However, it would appear that a good deal of the time for implementation of the model comes from bookkeeping.

A coarser model might lump bit strings at a grid of vaccine-antibody, vaccine-antigen distances together and discretize time so that event rates are kept constant over fixed stretches. Then the discretized process is no longer Markov but hidden Markov. To generate it correctly one would in fact have to go to the original computer model and simply record less. However, if the lumping of states and discretization of time is not too extreme one might hope to approximate the generation of events by a Poisson vector with parameters depending on the state as described in coarsened form at the previous discretized time point. The transfer matrix would then combine the Poisson vector and previous state to give the current state

4. SUMMARY OF DISCUSSIONS

Following each session there was a lively open discussion led by a moderator and summarized by a rapporteur. These discussions raised many issues but provided few answers, nevertheless, they were immensely valuable. They stimulated thinking, raised awareness, inspired collaborations, and helped identify directions for future research. The following provides a summary of these discussions and the rapporteurs' syntheses of the discussions framed by the four questions. While this section captures some of the content of the discussions, it does not adequately convey their energy or enthusiasm. The following summary is compiled from notes, rapporteur vu graphs, tapes and reviewer comments; as such it does not necessarily reflect the opinions of the authors.

4.1 What do we mean by model evaluation?

This question met with the resounding response that to determine how good a computer model is, one must first answer the question "good for what?" Indeed, a clear message from the workshop was that there is really no such thing as a context-free model evaluation. A second message was that, while context is key, it is too often not taken seriously. A client's challenge, such as "what is this model good for?" may be required before the modeler begins to appreciate the ways in which context might matter.

During the discussions, components of context were identified including:

- Application. Applications include science, where the aims of the model are exploration and explanation; decision–making, where planning and policy are emphasized; and training, where the computer model simulates real world situations to which individuals respond.
- Consequence. Consequence refers to the real-world impact of the modeling, such as impacts on public health, traffic, flood control, or safety as described in the examples in Section 3.
- Accuracy. Accuracy refers to the precision needed for the task at hand. It is one thing, for instance, to qualitatively represent the gross behavior of a wildfire and quite another to use that model to tell fire fighters when to abandon a position.

None of the workshop examples included a training application (although one of the future goals of the wildfire control model was to be able to use it to help train fire fighters), but all have science and decision-making contexts. The application context for Mesoscale Model Version 5 (MM5) in the workshop example was to provide input to a hydrological model to predict streamflow and runoff in support of decision-making to mitigate flood disasters. The evaluation function was defined in terms of predicting spatially distributed 48-hour precipitation accumulation, the input to the hydrology model. What needed to be weighed were the consequences of error for not acting to prevent flooding when necessary, versus taking (costly) actions when not necessary. The probability of decision errors had to be small, thus requiring great accuracy for the input to the hydrology models. Output from two simulations of MM5 (in principle, two different models, one using Four Dimensional Data Assimilation (FDDA)), were compared with data from a single storm event. Neither simulation produced the requisite accuracy. However, comparison of the two simulations revealed important differences between them, illuminating the science even though the decision requirements could not be met.

The traffic model, CORSIM, was used both for scientific exploration (what factors are important when designing traffic signal timing plans?) and ultimately to support decision-making (which of a number of possible signal control strategies

should be implemented?) The consequences of errors for the decision-making application ranged from irate phone calls to the traffic chief to major traffic congestion. An interesting result of the model evaluation was that when the model failed (unreasonably long queue times and unrealistic spillback), the failure led to insights about traffic modeling: rolling stops rather than complete stops are needed to capture driving behavior. This result supported the point made several times during the workshop that an incorrect model can still be useful. Moreover, it was noted in the discussion that one might be able to make a good decision, even if one could not make an accurate prediction.[2]

The immediate goal of the wildfire model FIRETEC was to facilitate better decision-making. A future goal was for the model to help guide the actions of fire fighters in real time (e.g., "evacuate an area now"). The consequences of modeling would then be serious indeed. Interestingly, the modelers aspired to these ambitious decision-making goals even though there was apparently no hope of capturing the exact behavior of any single fire.

The influenza model application context ranged from the scientific concern of "does the model explain the phenomenon" to the policy concern of "should an individual be vaccinated each year?" The discussion raised again the prospect of gaining scientific understanding and/or making useful predictions even with models of the "emergent" or "phenomenological" type (in the terminology of Section 2).

4.2 What makes model evaluation difficult?

For this question there was no shortage of answers. One answer, common to all of the models, was "the available data are inadequate." Either the data were poor (rain gauge data poorly positioned and crudely measured) or limited (wind tunnel experiments or controlled burns instead of real fires). It appeared to some that model evaluation produced a difficult double-bind. Models are necessary to predict/understand phenomena when there are very limited or no data. But without data, model evaluation is extremely difficult. So, when you need models the most, they are least likely to be reliable. In addition, the statisticians in the audience (and perhaps others) felt that the modelers did not sufficiently emphasize or plan for the collection of data on which model evaluation might be based. There was also concern that modelers were not aware of, or at least had not considered using, rich new sources of data. A few participants even wondered if in some instances the modelers were ignoring the data that might answer the scientific or policy question better than any of the existing computer simulations.

---

[2] Subsequent to the workshop, a follow-up study was done on a larger, more complex network near downtown Chicago. Data were collected in May, 2000 and a validation exercise similar to the one discussed in 3.3 was carried out. Flaws in the model were detected and adjusted for in devising a new signal plan and the recommended plan was implemented by the Chicago Department of Transportation in September, 2000. Data were collected that confirmed the prediction that the new signal plan would improve traffic flow ("predictive validity").

At the same time, for many applications raw data as such does not really exist. It comes having already been massaged by various physical devices, computer algorithms, and statistical computations. Thus a comparison between simulation output and data, for example, is really a comparison of two amalgamations differing in the amount and use of external information. Moreover, the data are often not observations of the physical variables needed for comparison with the model. In these cases, there is a secondary model that takes the raw data and converts it to the quantities that will be used in the model evaluation. In short, the data used in model evaluation are far from pristine and this adds another layer of uncertainty and difficulty to the model evaluation process.

Other issues identified as making model evaluation difficult included the following:

- There are qualitative measures as well as quantitative measures for model evaluation. In some cases, what most interests a modeler may be emergent qualitative behaviors, such as the Gulf Stream or traffic jams or qualitative agreement with epidemic data.
- Data assimilation is difficult and takes many forms, some of which do not have sound statistical foundations. The criteria used to evaluate a model with ongoing data assimilation may be different from those for a model used to make forecasts without such information. Indeed, it may make little sense to use the same data for model corrections and later, model evaluations.
- Calibration is difficult given the large number of variables involved. If calibration has to be done on the same data used for validation (as is often the case in statistics), then the evaluation becomes model dependent and less credible (unless account is taken of the dual use).
- Propagating uncertainty is difficult because there are numerous and often poorly understood sources of uncertainty: in the science, the translation of the science into computer code, and in the numerics, software bugs and hardware, as well as data.
- The size, complexity, and resource requirements are often immense so that multiple runs or ensembles of runs may not be practical.
- There is no microtheory, e.g., no Navier-Stokes equation, underlying many models.

4.3 What strategies can be employed?

The model evaluation strategies identified in the workshop (only some of which were actually used in the examples) included the following:

- *Comparing model output to real-world data.* Stochastic CORSIM simulations were compared with observed queue lengths; MM5

precipitation forecasts were compared with 48-hour averages from rain gauge data. Sometimes these comparisons were qualitative, that is, feature comparisons. For example, FIRETEC simulations showed wind gusts, fire-line shape and areas not completely burned, which were qualitatively similar to those observed in real wildfires and the immunology model predictions were borne out.

- *Comparing model output to experiments.* FIRETEC simulation output was compared to output from wind tunnel experiments and to controlled burns, where at least some of the site-specific conditions prior to the burn had been measured.
- *Comparing different models.* In Fovell's presentation, MM5 output without data assimilation was compared to model output using data assimilation. Perelson suggested developing independent implementations of the antigenic distance model and comparing results.
- *Comparing model to theory.* For example, the direction of flow of the main Atlantic and Pacific gyres must agree with the laws of physics governing fluids in a rotating system. This type of comparison is done more or less formally throughout model development and application.
- *Comparing model performance to experience/expert judgement.* Spillback appeared to be too great in traffic simulation, or queue times were too long in some runs. These comparisons are also done more or less formally for most models during development and application.
- *Varying resolution and physics* (including the parameterization of unresolved effects). This strategy was followed by LANL ocean modelers to determine the resolution required to resolve narrow boundary currents and mesoscale eddies.
- *Comparing model output distributions across many runs* (statistical summaries) with data. For stochastic models such as CORSIM or the antigenic distance model, as well as for deterministic models for which input and boundary conditions are only statistically known, both the mean and the spread among model runs need to be compared with the available data.

4.4 What is the role for statistical concepts and tools, and where are the statistical gaps?

Not too surprisingly there were lots of ideas for the role of statistical concepts and tools. Gaps can be inferred from answers to the "what makes evaluation difficult" question, and some of the recurring "gap" themes are listed below.

Available tools include the following:

- Statistical techniques for sensitivity and uncertainty analysis. These continue to be a subject of very active research. Some new directions have recently been proposed by Kennedy and O'Hagan (2000).
- Statistical approximations to numerical code. Statistically equivalent models (SEMs) were mentioned by Schoenberg in the context of FIRETEC for assisting with inversion or sensitivity analysis of a model, and by Bickel as an alternative implementation of the antigenic distance hypothesis. They may also replace physical submodels within a large, complex computer model.
- Outlier analysis for detecting bugs and other anomalies.
- Various statistical tools for calibrating and improving models including combining data and model (e.g., nudging and data assimilation).
- Developing functionals of model output, which optimize discrimination among models or model parameterizations relative to given criteria.
- Statistical graphical techniques for visualization and statistical pattern recognition and imaging techniques to quantify differences in qualitative features.
- Experimental design for computer experiments and data collection, particularly in the presence of major practical constraints. Modeling the contribution of the uncertainty (biases) introduced by such constraints.
- Statistical techniques for tuning input parameters (model calibration).
- Adjoint methods, in particular using statistically equivalent models for the efficient approximation of adjoints and to quantify sensitivities to inputs.
- Statistical techniques to reduce dimensionality (principal components, other orthogonal decompositions such as wavelets, etc.)
- Analysis of covariance structure of observational data versus that of model predictions.

Some of the "gap" themes identified included bringing formalism to "eye ball" or "viewgraph" comparisons, and to statements such as "the simulation and real world pictures of the Gulf Stream *look alike*." A theory of multiscale phenomena—in particular techniques to quantify uncertainty when moving between scales (e.g. moving from finer local models to larger-scale models)—is needed. Better definitions and theory are needed for what is meant by *model uncertainty*. There is also a need for more clarity about the differences between model calibration and model evaluation, particularly when some of the available data are used for calibration. Better tools are needed for comparing competing models and discriminating between models.

In addition to these gaps, Greg McRae (MIT) identified the need to build deterministic models that deal with model uncertainty from the beginning. He maintained that "if uncertainty is crucial to the model, then putting it in the beginning is critical." McRae asserted that this approach "will lead to very different models" and that "physical scientists, applied mathematicians and statisticians must work together to do it."

4.6 Additional Issues

While the discussion was rather focused on the four initial questions posed by the workshop organizers, a number of other related issues surfaced and were briefly discussed. These included the following.

- Detail vs. accuracy: adding more detail to a model (e.g, "more physics") is often expected to provide better model predictions. However, Rod Linn (wildfire control modeler) and others noted that in their experience more detailed modeling does not necessarily give better results. Linn explained that adding more detail to the models was not a good thing if the basis for the inclusion of the detail was poorly understood, incomplete, or incorrect. Linn noted that for the wildfire model adding more physics often led to worse results.
- Feedback between model construction, prediction and data analysis could improve models. A key obstacle is that, frequently, different people work on the research in its various phases. One obvious step toward a solution is to bring all the players on board from the beginning.
- Simulation flaws can teach us about the real world. An example given above was the discovery of the unmodeled driver behavior in the CORSIM example.
- Statistical tools may not be useful before the model reaches a certain stage of development. Primitive versions of models (ocean models for example) are "wrong by inspection". On the other hand, it was suggested that the role of statisticians at the front end of model development should be encouraged. Is the difference between these two points of view one of context? Might statisticians be particularly useful when the model is being developed to aid in decision-making, as opposed to when pure research is the aim?

5. WORKSHOP CONCLUSIONS AND NEXT STEPS

The workshop generated substantially more questions than answers, however it did much to focus the research issues. Many of these have been alluded to in the article, the introduction, the examples, and the responses to the four questions. To

summarize, computer model evaluation will benefit from research and the development of methods and tools in a number of areas. None of these areas is the exclusive domain of statisticians, but statisticians have something to contribute to all of them, particularly when working in collaboration with modelers, applied mathematicians, and other scientists.

- Data for model evaluation: How do we collect data to better evaluate computer models? For example, how might we use the results of adjoint or other sensitivity analysis methods to determine what sorts of data would be of most help in model evaluation? In general, however, models are likely to continue to be evaluated using data collected for other purposes, or data limited by technological and practical constraints. How do we formally characterize the effects of biases or constraints in such data, when used to evaluate models or estimate model uncertainty?

- Computer experiments: How do we design and implement computer experiments, especially when computation is costly? In particular, how do we design to get the information that may most effectively be played off against available data or the output from competitive computer models?

- Comparing model output and data: How do we deal with the differences between the time and space scales provided by model output and those inherent in the available data? How do we optimize functionals of both output and data to maximize discrimination between model and data, or between models? What key features do you extract from very high dimensional model output or data? And finally, how do we account for the additional uncertainty in model evaluation introduced by the necessity to manipulate data and/or model output in order to make them comparable?

- Statistically Equivalent Models (SEMs): How do you develop good statistical approximations of computer code (parts of models or entire models) to run as substitutes for the computer code? Many statistical tools that one might like to use for model evaluation will not work on the computer code, for computational or structural reasons, but will work on SEMs. Many non-statistical approaches to "metamodel" development have also been proposed (neural nets, fuzzy logic, etc.); should statisticians incorporate these options into their standard repertoire?

- Competitive Statistical Models (CSMs): Can we make better use of statistical models developed inductively by fitting data? If such statistical models do a better job of fitting the data and forecasting, they indicate not only that the computer model is not using all the information available, but often what information is not being properly exploited. Can we identify submodels that can be replaced by statistical models without loss of

overall model performance, possibly only as a temporary expedient, so that more computationally intensive model evaluation can be carried out?

- Sensitivity analysis: What statistical methods compete with the adjoint methods available for computer models based on differential equations? How do we optimize statistical methods for models for which such methods are unavailable?

- Uncertainty analysis: The sources of uncertainty in computer models are so varied and complex that existing tools are not yet fully up to the task. Different tools are probably needed for different sources of uncertainty and different model applications. How do we quantify the effects of structural uncertainty in the model or the underlying theory? Of alternative parameterizations of unresolved features and effects in model? Of numerical error and implementation errors (bugs)?

In a few of the areas listed above, there is already useful work under way, and indeed a history of useful work. Computer experiments is perhaps the best illustration (e.g., McKay et al. 1979, Sacks et al. 1987, and much recent work.) In those cases, the workshop message is do more and do more in close collaboration with modelers. For the other areas, there is either the need for the transfer of statistical methods and tools to new applications or the need for new statistical methods and tools. A good illustration is in the comparisons between large, multidimensional datasets and large, multidimensional computer output. Statisticians can already bring a lot to the table, but in addition, there are a number of difficult and novel problems that need to be solved that may be related, for instance, to current work on imaging and uncertainty analysis. In addition, there is a need for software tools and and other methods for disseminating statistical approaches to the modeling community.

The final question was how do we get started addressing these questions? Many good suggestions were made in the concluding session of the workshop, including the following.

- Develop a website for communication of the participants in this workshop and others interested in continuing the discussions.
- Conduct smaller workshops on specific aspects of model evaluation.
- Promote education among statisticians: probability and statistics that use real work examples for the modelers as well as science and engineering for the statisticians and mathematicians.
- Start collaborations, and develop an understanding of why some collaborations work better than others.

- Maintain and expand training opportunities for graduate students and post docs in statistics in settings like the national laboratories and the National Center for Atmospheric Research where computer model evaluation problems in multidisciplinary settings abound.
- Work to break down funding barriers for collaboration, e.g., modelers are reluctant to spend scare resources on statistical assistance, particularly when the worth of such collaborations has not yet been widely established.
- Support statisticians who can promote the computer model evaluation area at high levels within funding agencies.
- Develop a framework and consistent language to address evaluation techniques that the modeling community could use.
- Develop new and useful research in the area.

Finally, the most notable recommendation was that we **do something** to keep the energy and momentum generated during the workshop moving such that we begin to resolve the many and varied problems related to computer model evaluation. We hope that this article, the open questions raised, and the list of continued activities will keep us moving forward.

6. REFERENCES

BAYARRI, S. and BERGER, J. (1999). Quantifying Surprise in the Data and Model Verification. In *Bayesian Statistics* 6, (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.) 53-82. Oxford University Press, London.

BEYER, W. E. P., DEBRUIJN, I. A. (1998). WESTENDORP, R. G. J. and OSTERHAUS, A. D. M. E. *Vaccine* **16**, 1929-1932.

BOSSERT, J., REISNER, J. M., LINN R. R., WINTERKAMP J. L., SCHAUB R., and RIGGAN, P. J. (1998). Validation of Coupled Atmosphere-Fire Behavior Models. Preprints of the 14th Conference on Fire and Firest Meteorology, November 16-20, 1998, Luso-Coimbra, Portugal.

BOX GP and DRAPER N. (1987) Empirical Model Building and Response Surfaces. J. Wiley, New York .

CLARK, T. L., RADKE L., COEN J., and MIDDLETON D., 1999: Analysis of small-scale convective dynamics in a crown fire using infrared video camera imagery. J Appl. Meteor., 38, (in Press)

CLICK, S. and ROUPHAIL, N. (1999). Lane Group Level Field Evaluation of Computer-based Signal Timing Models, Paper presented at the 78th Annual Meeting, TRB.

CORSIM USER'S MANUAL. (1997). FHWA, U.S. Department of Transportation, Office of Safety and Traffic Operation R&D, Intelligent Systems and Technology Division, McLean, VA.

EFRON B. and TIBSHIRANI R. (1993). Introduction to the bootstrap. Chapman and Hall, London.

GOLDBERG, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading.

HOSKINS, T. W., DAVIS, J. R., SMITH, A. J., MILLER, C. L. and Allchin, A. (1979) *Lancet* **i**, 33-35.

KIEFER J. and WOLFOWITZ J. (1952) Stochastic estimation of the maximum in a regression. Ann. Math. Stat. 462-466.

KEITEL, W. A., CATE, T. R., COUCH, R. B., HUGGINS, L. L. and HESS, K. R. (1997) *Vaccine* **15**, 1114-1122.

KENNEDY, M. AND O'HAGAN, A. (2000). Bayesian Calibration Of Computer Models. http://www.shef.ac.uk/~st1ao/pub.html. To be published in the Journal of the Royal Statistical Society, Series B.

KUSHNER H., and LIN G. (1997). Stochastic Approximation algorithms and Applications. Springer-Verlag Inc., New York .

LINN, R. R. (1997). Transport Model for Prediction of Wildfire Behavior, Los Alamos National Laboratory Scientific Report: LA13334-T.

MCKAY, M. D., BECKMAN, R. J. and CONOVER, W. J. (1979). Comparison of three methods for selecting input variables in the analysis of output from a computer code. Technometrics 21, 239-245.

MILLER, R. G. (1966). Simultaneous statistical inference. McGraw Hill, New York

PERELSON, A. S. and OSTER, G. F. (1979). *J. Theor. Biol.* **81**, 645-670.

SACKS, J., SCHILLER, S. B. and WELCH, W. J. (1989). Designs for computer experiments (with discussion). Technometrics 31, 41-47

SCHOENBERG, F., BERK, R., FOVELL, R., LI, C., LU, R., and WEISS, R. (2001). Approximation and Inversion of a Complex Meteorological System via Local Linear Filters. J. Appl. Meteorology, to appear.

SMITH, D. J., FORREST, S., ACKLEY, D. H. and PERELSON, A. S. (1999). Proc. Natl. Acad. Sci. USA **96**, 14001-14006.

SMITH, D. J., FOREST, S., ACKLEY, D. H. and A. S. PERELSON (2000). Variable efficacy of repeated annual influenza vaccination, PNAS. 96 14001-14006.

STORCH, H. von and NARVARRA, A. (1999). Analysis of Climate Variability: Applications and Statistical Techniques, Springer-Verlag.

ZEIGLER, B. (1976). Theory of modelling and simulation, John-Wiley, New York.