Title
Some Statistical Tools for Evaluating Computer Simulations: A Data Analysis Approach

Authors
Richard A. Berk
Robert G. Fovell
Frederic Schoenberg
et al.

Publication Date
2011-10-25

The Use of Statistical Tools for Evaluating Computer Simulations*

Richard A. Berk
Department of Statistics

Robert G. Fovell
Department of Atmospheric Sciences

Frederic Schoenberg
Department of Statistics

Robert E. Weiss
Department of Biostatistics

May 29, 2001

*The work reported in this paper was supported by a grant from the National Science Foundation, Program in Integrated Assessment.

1 Introduction

Climate scientists who work with computer simulation models readily appreciate that the quality of their science depends upon the quality of their simulations. Good modeling may be recognized, in part, by the serious application of various assessment tools to help determine the credibility of model output (Hänninen, 1995; Skiles, 1995; Shackley et al., 1998; Grassl, 2000). As instructive as these efforts are, however, we argue below that climate scientists typically fail to properly exploit a number of useful statistical procedures. Moreover, their assessment efforts are too often ad hoc and lacking in much formal justification. Modelers also typically do not appreciate that the assessment tools are only tools. Their effective use depends upon a general strategy of model evaluation embedded more thoroughly in an epistemology underlying much of modern statistics. Finally, we briefly describe several important classes of model evaluation problems for which no good solutions exist and for which a concerted research effort is needed.

2 An Overview of Current Model Evaluation Practice

It is common, though hardly universal, for climate scientists to "validate" their computer models in various ways.[1] Sometimes simulation output is compared to empirical observations or to expectations from accepted theory. If the simulation output is badly inconsistent with either, model revisions are typically implied. For example, the output of general circulation models (GCMs) can be compared to data from ice cores characterizing climate variation over many thousands of years in the past. If well documented patterns of warming and cooling are not reproduced, the GCM may need to be substantially revised (Barron, 1995; Crowley and Berner, 2001). There sometimes also are forecasting assessments, such as recently conducted for computer models of the El Niño phenomenon (Syu and Neelin, 2000b; Kerr, 2000). Weak forecasting performance can often mean problems with the computer model. Yet another approach is to compare intermediate and final model output to the results of laboratory experiments (Lu et al., 1997a).

[1] It may be more than a quibble to point out that climate scientists "validate" models while statisticians "evaluate" models. In principle, the tools applied should be the same, but clearly the aspirations are quite different.

Finally, computer experiments often are undertaken in which input data are perturbed or in which small parts of the computer code are manipulated to see whether the model is overly sensitive to such changes. Such studies are common in regional scale modeling of climate impacts (Bonan, 1997) and elsewhere (Syu and Neelin, 2000a; Moorthi, 2000).

2.1 The Use of Summary Statistics

In validation efforts such as these, statistical tools are sometimes used. For example, it is common to compute one or more descriptive measures of "error." Lu and his colleagues (1997b), for instance, compute (among other things) the "root-mean-square difference" between simulation output from their regional air quality model and measured values from local air quality monitoring stations. Mearns and her colleagues (1997) use as a summary measure the ratio of two crop yield means, one based on data and one based on a climate simulation model.

Summary measures can be useful, and indeed are essential, given the large volume of computer output commonly examined. However, many popular summary measures throw out enormous amounts of information and as such are at best a start in evaluating a model's performance. If the simulation output is arrayed in a three-dimensional spatial grid, for instance, a single summary measure such as root-mean-square error provides no indication whether the model reproduces the data better in some locations rather than others. Visually inspecting the output and data on comparable maps can be a step in the right direction, but one risks turning the exercise into a kind of beauty contest in which the conclusions are solely in the eye of the beholder. As a perceptual matter, is also extremely difficult to make level or ratio comparisons between locations separated by even modest distances (Cleveland, 1993). Soong and Kim (1996), for example, simulate a "heavy wintertime precipitation event" in California, and supplement their summary measure of fit (a Pearson correlation) between the simulation output and measured precipitation with a comparison between maps of measured precipitation and simulated precipitation. They conclude that the maps look much alike, but it may have also been useful if they could have been more expansive and precise. The general lesson is that the popular summary statistics which attempt to encapsulate enormous amounts of information may obscure more than they enlighten.

One must also wonder about how well the underlying properties of pop-

ular summary statistics sometimes are appreciated. To take a very simple example, root-mean-square-error depends on a quadratic loss function implying a) that larger deviations between the data and the output are given especially heavy weight in the computations and b) that "overpredictions" have the same scientific import as "underpredictions." The former means that the root-mean-square error is being driven by the tails of the disparity distribution, and outliers can dominate the results. The latter means that simulation values that are too large are treated the same as simulation values that are too small. If these consequences comport well with the scientific issues at hand, all is well. If they do not, misleading conclusions can follow. (And there are usually a number of alternative procedures.) The point is that the underlying properties of summary statistics need to be carefully considered. For popular summaries of multivariate relationships, such as regression analysis and principal components analysis, this can be a demanding exercise even though over the past two decades, there have been great advances in diagnostic tools that can help enormously (e.g., Cook, 1998a). Unfortunately, for a wide variety of recent computer intensive procedures, such as those popular in data mining, many performance characteristics are not even well understood.

Finally, the data used as "ground truth" are almost never what the name implies. There are commonly important biases and noise. In addition, most data are heavily processed before they are used. Indeed, the data are often derived from an auxiliary set of data processing computer models whose impact can be dramatic. Sometimes, and perhaps unavoidably, the models used in that processing may be same ones that created the output requiring evaluation in the first place. As an example, in compensation for instrument errors and/or gaps in the observational network, the "ground truth" data set may be constructed using a first-guess field provided by the same forecast model that is being assessed. Even if two nominally independent models are employed, however, the independence has can questioned because as some have claimed, "climate or forecast models are more like each other than they are like the atmosphere" (Daley and Mayer 1986).

Such complications are certainly not news to climate scientists, but there seems to be too little appreciation of the fact that summary statistics can be badly distorted and that sometimes these statistics can be altered to better respond the real properties of the data. For example, if noise is a significant problem and if the variance of the noise can be estimated, correlation coefficients can be corrected for attenuation.

2.2 The Use of Statistical Inference

It is also common to apply formal statistical inference in model evaluations. For example, Mearns, Rosensweig and Goldberg (1997) consider simulated crop yields under different climate change scenarios over nearly 100 years, and use a Kolmogorov-Smirnov 2-sample test to evaluate differences between the cumulative distributions of yields (over years) from the different simulations. Confidence intervals are also popular. Williams, Shaw and Mendelsohn (1998), for instance, study the impact of climate changes on the value of land used for farming. They supplement conventional tests on differences between means for GCM-based predictions and data-based estimates, with bootstrap 95% confidence intervals to calibrate average differences between the two.

In certain situations, tests and confidence intervals can be very instructive, but they may be misleading as well. One key assumption usually is independence. For example, each output value from the simulation model is assumed to be independent of every other output value. This usually is untrue on its face. If tests are undertaken nevertheless, one risks being seduced by falsely high precision, and the null hypothesis may well be rejected when it should not have been. Ideally, one would like to take any dependence properly into account before such tests are implemented. There is also the deeper question of what it means to apply statistical tests to deterministic models; there is no well-defined chance mechanism whose impact needs to be assessed and as a result, one is well down the slippery slope toward statistical ritual (Berk and Freedman, 1995, Berk, 1995).

This point warrants some elaboration. With conventional "frequentist" statistical inference, the estimated probabilities are based on a thought experiment in which the data set is generated a limitless number of times by a chance mechanism, with every replicate generated independently of every other. For deterministic computer simulation models, there is a mismatch between this thought experiment and how the data were actually produced. If one ran the models over and over again, the output would be effectively identical. There is no stochastic component.

One can certainly introduce variation into the outputs by changing the inputs to each simulation or by changing the computer code before each simulation began. Indeed, this is often done by climate scientists under the rubric of sensitivity analysis. But there are at least two serious problems.

First, in order to justify formal statistical inference, these changes would have to be produced by a well understood and implemented *chance mech-*

anism, which is usually not the case in sensitivity analysis (e.g., Moorthi, 2000). It cannot be overemphasized that altering inputs or model features in a fixed manner (i.e., without a chance mechanism) does not suffice; the outputs formally are *still* deterministic and frequentist tests *still* do not apply.

Second, even if chance variation can be introduced for comparisons between models or between a model and data, analyses based on that variation miss the point; the question at hand is to characterize any *existing* disparities. To introduce a new source of error, even by a chance mechanism, would lead to tests that do not address the original comparisons of interest. That is, a problem is created that can be solved rather then solving the real problem at hand.

Of course, one could argue that the introduced uncertainty is an attempt to represent chance processes that are actually present in the empirical world, but lost in the deterministic simulation code. If so, one must wonder why chance processes were not built into the code to begin with. Moreover, this view requires that one have good theory and empirical support for the chance mechanism proposed. Otherwise, the simulation risks characterizing some science fiction world, not the one science cares about. Can a modeler really make the case, for instance, that random variation in each of several initializing conditions behaves as if drawn independently and at random from a particular joint distribution, with constant parameters (e.g., means and variances)? Note that there are perhaps a half dozen assumptions implied, all of which need to be at least approximately true for the actual phenomena being modeled. We are skeptical that in most cases such assumptions could withstand much scrutiny (Berk et al., 1995; Berk and Freedman, 1995). For example, the empirical observations used to initialize GCMs are not readily conceptualized as proper random variables from some joint distribution, in part because they share the same difficulties as ground truth data and in part because they are typically not generated by well understood chance mechanisms.

It is also possible to reformulate the problem so that the model is properly treated as deterministic, and uncertainty is introduced by the ground truth data alone. The summary statistics from the simulation are taken as fixed population values. And given those population values, coupled with certain assumptions, one can construct the probability density or distribution of the summary statistics from the data (i.e., the sampling distribution) under a limitless number of independent "trials" by which the data could in principle be generated. Conventional tests and confidence intervals then

naturally follow. However this assumes that the *data* are generated by a well understood chance mechanism (e.g., probability sampling). As noted above, ground truth data commonly have extremely complex properties not well captured in a frequentist thought experiment. More likely, the data are a population or a convenience sample. In the first case there is no uncertainty, and the thought experiment of limitless independent trials does not apply. In the second case, there is typically no way properly to represent the chance process by which the data were produced (Berk and Freedman, 1995). Then, conventional tests and confidence intervals do not apply.

To further complicate matters, there are other kinds of uncertainty linked to such things as how well the underlying physics is represented in the computer code or to the values of certain parameters. These cannot easily be placed in a frequentist framework. The Bayesian alternative, based on subjective probabilities, could in principle be applied, but uncertainty that comes from potential problems with the computer model itself are not well handled by Bayesian approaches either, unless the computer model is reformulated from the ground up. For example, one could in principle place a prior joint probability distribution on all of the model's parameters, but this is not usually the way simulation models are conceptualized and would be a Herculean task requiring yet another set of assumptions.

2.3 What do the Experts Say?

The broad set of problems described above are not aberrations. A recent collection of didactic papers on the analysis of climate variability includes three chapters addressing model validation, all of which rely heavily on overall summary statistics and hypothesis tests. Frankignoul (1999) argues for the use of conventional t-tests and its multivariate generalizations. Briffa (1999) makes the case for bivariate Pearson correlations (and associated statistical tests) and eyeball inspection. Livezey (1999) considers resampling procedures borrowed from statistics for situations in which conventional distributional assumptions cannot be met. These didactic expositions of how model assessment should be done would seem to encourage many of the difficulties just described. It is not surprising, therefore, that even when climate scientists try to do the right thing, the model evaluations can be unsatisfying (e.g., Greve, 2000; Van Minnen et al., 2000, section 4.2).

Even more to the point are the conclusions from a recent workshop on evaluating complex computer models held at the Los Alamos National Lab-

oratories, sponsored by the the National Research Council's Committee on Applied and Theoretical Statistics and by the National Institute of Statistical Science.[2] Perhaps the key conclusion was that model evaluation is a daunting enterprise and solutions do not yet exist for a variety of important problems. Specific difficulties include a lack of adequate "ground truth" data, characterizing and propagating uncertainty, and making more than superficial comparisons between large, high-dimensional output and the available data. There was also agreement, however, that whether because of "culture" or the incentives that modelers face, conventional model evaluation practice typically fails to exploit a host of modern strategies and procedures from statistics and applied mathematics (Berk et al., 2000a). Sophisticated statistical tools are certainly appreciated and applied in climate science (Daley, 1993; Xue et al., 1994; Jiang et al., 1995; Wilks, 1995; Ghil and Yiou, 1996). However, these techniques are used primarily in empirical, data-driven studies, not in computer model evaluations.

3 Potential Statistical Contributions

Most of the work that has appeared in statistical journals addressing computer simulation models focuses on the sensitivity of the simulation output to variation in simulation input. Some work considers methods to decompose variation in model output as a function of simulation input (Saltelli et al., 1999; Archer et al., 1998). In principle, one can then learn which inputs are most important in driving model outputs. Related other work examines how best to develop statistical models of input/output relationships (Currin, et al., 1991; Haylock and O'Hagan, 1996; Lim et al., 1997; Kennedy et al., 1999; Schoenberg et al., 2000), often with the intent of designing sets of computer simulation experiments so that one is able to make good use of each computer run (Sacks et al., 1989). Computer simulations can be very costly. The strategy, therefore, is to choose in a highly informed manner which computer simulations to run.

At the same time, there are a range of existing statistical tools that could be easily transferred to model evaluation problems. Obvious examples are the tools used in data analysis. One treats the computer inputs and outputs as a virtual world that needs to be described in much the same way that one

[2]Over 100 modelers, mathematicians, and statisticians participated in the two-day workshop.

would try to describe the relationships between measured variables in the real world. When there are "ground-truth" data as well, their relationships to the virtual output data are another essential analysis. An effective description of the virtual world would provide insights into *how* the computer simulation is performing, which in our view is a necessary step before deciding what about the model is satisfactory and what about the model is not. The "how" might be summarized in a few simple statistics, but more likely would require at least several different kinds of summaries addressing different features of the simulation. In addition, statistical graphics are likely to be essential because patterns in the data, not easily represented by numerical summaries alone, can sometimes be critical.

From this perspective, one could draw on the long-standing tradition of exploratory data analysis (Mosteller and Tukey, 1977; Cleveland, 1993) which, thanks to modern computing, is enjoying remarkable growth. Examples include sophisticated smoothers (Chambers et al., 1979; Hastie and Tibshirani, 1990), graphical regression (Cook, 1998a), data mining (Breiman et al., 1993; Hand, 1998), and dimension reduction procedures such as sliced inverse regression (Duan and Li, 1991) and principal Hessian directions (Cook, 1998b). However, the goal remains the same: to exploit as much of the available information as possible to gain understanding about what may be going on. While for some this is merely a "blue-collar" enterprise, it is no less essential than in scientific work more generally.

In addition to statistical tools and strategies that might be effectively transferred, there are a host of unsolved model evaluation problems on which modelers, applied statisticians, and statistician might jointly make important contributions. Building on the Santa Fe workshop noted earlier (Berk et al., 2000a), we list very briefly below some key issues. Readers interested in the technical details are urged to examine the workshop report.

- *Data for Model Evaluation* — How does one collect better data for model evaluation or better assess the properties of existing data? For example, how might adjoint methods (Talagrand and Courtier, 1987) be used to design formal sampling plans taking into account especially important locations in time and space. There are several traditions in statistics that might be exploited, such as adaptive sampling (Thompson and Seber, 1996) and model-based sampling (Cumberland, 1998). One advantage of such methods (and probability sampling more generally), is that uncertainty resulting from sampling error can be properly

represented. This is not true for "convenience samples" or "judgment samples," which are not collected using explicit chance mechanisms.

- *Computer Experiments* — How does one design and implement computer experiments when computation is costly? This is an area in which there has already been significant statistical work, as noted above. Matters are substantially more complicated when it is not clear in advance exactly what analyses will be undertaken (which is almost always the case). There are a number of new developments on model-robust designs (Li and Nachtsheim, 2000) that could well be useful.

- *Comparing Model Output to Data* — What features should one examine when making comparisons between very high dimensional model output and data? How then are systematic comparisons to be undertaken such that important scientific information is not obscured beneath layers of statistical manipulations? Current practice relies very heavily on eyeball assessments, but surely one can do better. For example, one might try to formalize what is done in good eyeball assessments building on insights from the statistical literature on imaging and vision (Ogden, 1997; Zhu, 1997; Simhadri et al., 1998). In addition, data mining and dimension reduction (noted above) could play a very important role.

- *Statistically Equivalent Models (SEMs)* — How can one develop statistical approximations of computer code to employ as substitutes for the computer code? Such approximation can bring with them a wide variety of benefits including a wealth of diagnostic tools that would otherwise be unavailable for model evaluation. One example is model checking plots (Cook, 1998a) in which model-based output and model-free output are formally compared. Insofar as the statistical model emulates key feature of the computer model, statistical diagnostics such as this can be used to make inferences about the computer model.

- *Competitive Statistical Models (CSMs)* — How can one make better use of statistical models developed to characterize the same empirical phenomena as computer models? If such statistical models do a better job fitting the data and/or forecasting, the computer model is not using all of the available information, or that information is not being properly exploited. Indeed, it may sometimes be possible for the statistical

model to suggest precisely where the computer model is inadequate (Berk et al., 2000b).

- *Sensitivity Analysis* — What statistical methods may be used in sensitivity analyses as alternatives to adjoint techniques when the statistical methods are more easily developed and/or provide more instructive results? For example, are there variations on, or generalizations of, the concept of influence (Cook, 1986) that might be instructively applied?

- *Uncertainty Analysis* — There are many different kinds of uncertainty linked to computer models, some associated with data used as inputs, with the parameterizations, with the level of resolution employed, with the nature of boundary conditions, with the numerical methods, with the basic science itself, and many more. Some of these are not easily conceptualized with conventional statistical methods, let alone combined in useful and practical ways. Clearly, there is important work to do characterizing uncertainty.

4 Conclusions

Few would dispute that a far better job could be done evaluating computer models used in climate science. The real issue here, therefore, is what role statistics can play. For many climate scientists, statistics may seem to be little more than applied mathematics. As such, the discipline of statistics at best offers some useful tools that can sometimes help in model evaluation. For some climate scientists, this provides enough motivation to import snippets of statistical technology. However, for statisticians the statistical procedures rest on a world view and set of related concepts that give the mathematics meaning. Absent an understanding of that world view and its related concepts, it is all to easy to overlook instructive statistical procedures, to use such procedures ineffectively, or simply to get it wrong.

At the same time however, statisticians hardly have all the answers. Indeed, for some of the most important and interesting problems, they have no answers at all. The good news is that a wonderful, joint research agenda for modelers, applied mathematicians, and statisticians is implied. But it will take a concerted and cooperative effort from all stakeholders to make significant progress.

5 References

Archer, G., Salteli, A. and Sobol, I.M. (1997) "Sensitivity Measures, ANOVA like Techniques, and the Use of Bootstrap," *Journal of Statistical Computation and Simulation*, 58, 99-120.

Barron, E.J. (1995) "Climate Models: How Reliable are their Predictions?," *Consequences: The Nature and implications of Environmental Change*, 1(3), 17-27

Berk, R.A., Bickel, P., Campbell, K., Keller-McNulty, S., and Kelly, E., and Sacks, J. (2000a) "Workshop on the Statistical Approaches for the Evaluation of Complex Computer Models, Working paper (under review), Statistics Group, Los Alamos National Laboratory, Los Alamos, New Mexico.

Berk, R.A., Bond, J., Lu, R., Turco, R. and Weiss, R.E. (2000b) "Computer simulations as experiments: using program evaluation tools to assess the validity of interventions in virtual worlds," in L. Bickman (ed.) *Research Design: Donald Campbell's Legacy*, pp 195-214, Newbury Park, CA: Sage Publications

Berk, R.A., Western, B. and Weiss, R. (1995) "Statistical Inference for Apparent Populations." In Marsden, P. (ed.) *Sociological Methodology, 1995*, pp 178-203, Cambridge, UK: Blackwell Publishing.

Berk, R.A., and Freedman, D. (1995) "Statistical Assumptions as Empirical Commitments." In Blomberg, T., and Cohen S. (eds.), *Law, Punishment and Social Control: Essays in Honor of Sheldon Messinger*: 245-248, New York: Aldine de Gruyter.

Bonan, G.B., (1997) "Effect of Land use on Climate in the United States," *Climate Change* 3, 449-486.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1993) *Classification and Regression Trees* New York: Chapman & Hall.

Briffa, K.R. (1999) "The Simulation of Weather Types in GCMs: A Regional Approach to Control Run Validation." In von Storch, H., and Navarra, A. (eds.), *Analysis of Climate Variability: Applications of Statistical Techniques*, New York: Springer.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, J. (1983) *Graphical Methods for Data Analysis*, Boston: Duxbury Press.

Cleveland, W.W. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74: 829-836.

Cleveland, W.S. (1993) *Visualizing Data*, Summit, New Jersey: Hobart Press.

Cook, R.D. (1986) "Assessment of local influence" (with discussion) *Journal of the Royal Statistical Society*, Series B, 48: 133-155.

Cook, R.D. (1998a) *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: John Wiley.

Cook, R.D. (1998b) "Principal hessian directions revisited" (with discussion). *Journal of the American Statistical Association*, 91, 84-100

Crowley, T.J. and Berner, R.A. (2001) " CO_2 and Climate Change." *Science*, 292: 870-872.

Cumberland, W.G. (1998) "Ratio, Regression, and Related Estimates in Sample Survey Methodology," in *Encyclopedia of Biostatistics*, P. Armitage and T. Colton (eds, New York: John Wiley.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991) "Bayesian prediction of Deterministic Functions, with Applications to Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, 86, 953-963.

Daley, R. (1993) *Atmospheric Data Analysis*. Cambridge: Cambridge University press.

Daley, R., and Mayer, T. (1986) "Estimates of global analysis error from the Global Weather Experiment Observational Network." *Monthly Weather Review*, 114: 1642-1653.

Duan, N., and Li, K.C. (1991) "Slicing Regression: A Link-Free Regression Method." *Annals of Statistics*, 19: 505-530.

Frankignoul, C. (1999) "Statistical Analysis of GCM Output." In von Storch, H., and Navarra, A. (eds.), *Analysis of Climate Variability: Applications of Statistical Techniques*, New York: Springer.

Ghil, M., and Yiou, P. (1996) "Spectral Methods: What They Can and Cannot Do for Climate Time Series." In Anderson, D.T.L., and Willebrand, J. (eds.) Decadal Climate Variability: Dynamics and Predictability. NATO ASI Series, Vol I 44, Berlin: Springer-Verlag.

Grassl, H. (2000) "Status and improvement of Coupled Circulation Models." *Science*, 288: 1991-1997.

Greve, R. (2000) "On the Response of the Greenland Ice Sheet to Greenhouse Climate Change," *climate Change*, 46, 289-303.

Hand, D.J. (1998) " Data mining: statistics and more?" *The American Statistician* 52: 112-118.

Hänninen, H. (1995), "Assessing Ecological Implications of Climate Change: Can We Rely on Our Simulation Models?," *Climate Change*, 31, 1-4.

Hastie, T.J., and Tibshirani, R.J. (1990) *Generalized Additive Models*, London: Chapman Hall.

Haylock, R. G. and O'Hagan, A. (1996) "On inference for outputs of computationally expensive algorithms with uncertainty on the inputs." In J.M. Bernardo et al (eds.) *Bayesian Statistics 5*: 629-637, London: Oxford University Press.

Jiang, N., Ghil, M., Neelin, J.D., (1995) "Forecasts of Equatorial Pacific SST Anomalies Using and Autoregressive Process Using Singular Spectrum Analysis," *Experimental Long-lead Forecast Bulletin* 4(1): 24-27.

Kennedy, M., Oakley, J., and O'Hagan, A. (forthcoming, 2000). "Uncertainty Analysis and Other Inference Tools for Complex Computer Codes." In J.M. Bernardo et al (eds.), *Bayesian Statistics 6*, London: Oxford University Press.

Kerr, R.A. (2000) "Second thoughts on Skill of El Niño Predictions," *Science*, 290, 257-258.

Li, W. and Nachtsheim, C.J. (2000) "Model-robust factorial designs," *Technometrics*, 42, 4, 345-352).

Lim, Y.B., Sacks, J, Studden, W.J., and Welch, W. J. (1997) "Design and Analysis of Computer Experiments When the Output is Highly Correlated Over the Input Space," Technical Report # 62, National Institute of Statistical Sciences, Research Triangle, North Carolina.

Livezey, R.E., (1999) "Field Intercomparison." In von Storch, H., and Navarra, A. (eds.), *Analysis of Climate Variability: Applications of Statistical Techniques*, New York: Springer.

Lu, R., Turco, R.P. and Jacobson, M.A. (1997a),"An Integrated air Pollution Modeling System for Urban and Regional Scales: 1. Simulations for SCAQS 1987," *Journal of Geophysical Research*, 102, 6063-6079.

Lu, R., Turco, R.P. and Jacobson, M.A. (1997b),"An Integrated air Pollution Modeling System for Urban and Regional Scales: 2. Simulations for SCAQS 1987," *Journal of Geophysical Research*, 102, 6081-6098.

Mearns, L.O., Rosensweig, C., and Goldberg, R. (1997, "Mean and Variance Change in Climate Scenarios: Methods, Agricultural Applications, and measures of Uncertainty," *Climate Change*, 35, 367-396.

Moorthi, S. (2000) "Application of Relaxed Arakawa-Schubert Cumulus Parameterization of the NCEP Climate Model: some Sensitivity Experiments," in Randall, D. A. (ed.) *General Circulation Model Development: Past Present and Future*, New York: Academic Press.

Mosteller, F., and Tukey, J. (1977) *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Ogden, R.T. (1997) *Essential Wavelets for Statistical Applications and Data Analysis*, Boston: Birkhäuser.

Sacks, J., Welch, W., Mitchell, T.J., and Wynn, H.P. (1989) "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409-435.

Saltelli, A., Tarantola, S., and Chan, K. P.-S. (1999) "A Quantitative Model-Independent Method for Global Sensitivity Analysis of Model Output," *Technometrics*, 41, 39-56.

Schoenberg, F., Berk, R.A., Fovell, R. G., Lu, R., and Weiss, R. E. (2000) "Approximation of Inversion of a Complex Meteorological System via Local Linear Filters," *Journal of Applied Meteorology*, forthcoming.

Shackley, S., Young, P., Parkinson, S., and Wynne, B., (1998) "Uncertainty, Complexity, and Concepts of Good Science in Climate Change Modeling: Are GCMs the Best Tools?," *Climate Change*, 38, 159-205.

Simhadri, K.K., Iyengar, S.S., Holyer, R.J., Lybanon, M., and Zachary, J.H. (1998) "Wavelet-based feature extraction from oceanographic images," *IEEE Transactions on Geoscience and Remove Sensing*, 36,3: 767-778

Skiles, J.W. (1995) "Modeling Climate Change in the Absence of Climate Change Data," *Climate Change*, 30, 1-6.

Soong, S.-T., and Kim, J. (1996) "Simulation of a Heavy Wintertime Precipitation event in California," *Climate Change*, 32, 55-77.

Syu, H.-H, and Neelin, J.D. (2000b) "ENSO in a Hybrid Model, Part I: Sensitivity to Physical parameterizations," *Climate Dynamics*, 16: 19-34.

Syu, H.-H, and Neelin, J.D. (2000b) "ENSO in a Hybrid Coupled Model. Part II: Prediction with Piggyback Data Assimilation," *Climate Dynamics*, 16: 35-48.

Talagrand, O., and Courtier, P (1987) " Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory," *Quarterly Journal of the Royal Meteorological Society*, 113 :1311-1328.

Thompson, S.K., and Seber, G.A.F. (1996) *Adaptive Sampling*, New York: John Wiley.

Van Minnen, J.G., Alcamo, J., and Haupt, W. (2000) "Deriving and Applying Response Surface Diagrams for Evaluating Climate Change Impacts on Crop Production," *Climate Change*, 46, 317-338.

Williams, L.J., Shaw, D., and Mendelsohn, R. (1998) "Evaluating GCM Output with Impact Models," *Climate Change* 39, 111-133.

Wilks, D.S., (1995) *Statistical Methods in the Atmospheric Sciences*. New York: Academic Press.

Xue Y., Cane M.A., Zebiak, S.E., and Blumenthal, B. (1994) "On the Prediction of ENSO: A Study with a Low-Order Markov Model," *Tellus* 46A: 512-528.

Zhu, S. C., Wu, Y., and Mumford, D. B. (1997)."Filter, random field, And maximum entropy (FRAME): towards a unified theory for texture modeling", *International Journal of Computer Vision*, 27,2, 107-126