Statistical Validation of Complex Computer Models

Rima Izem, Harvard University[1]

Introduction

This paper presents an overview of statistical validation of complex computer models. Such models—used, for example, to simulate traffic in a street network, a car crash, the effect of increasing CO_2 on global warming, or the cost-effectiveness of a procedures in cancer screening—play important roles in scientific research and policy and decision making. The computer models discussed in this paper are all based on mathematical models of the real phenomena of interest, as opposed to, for instance, models (such as those based on neural networks) that emulate the behavior of a phenomena without explicitly attempting to represent the underlying components.

Validating such models has been emphasized in at least two reports from the National Academies: NAS (1991) *"Improving Information for Social Policy Decisions -- The Uses of Microsimulation Modeling: Volume I, Review and Recommendations"* and NAS (1998) *"Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements."* Validation has also been the focus of a number of workshops. One notable example was a "Workshop on Statistical Approaches for the Evaluation of Complex Computer Models," held December 3-4, 1999 in Santa Fe, New Mexico as a joint activity of the Committee on Applied and Theoretical Statistics of the National Research Council, Los Alamos National Laboratory, and the National Institute of Statistical Sciences (NISS). Another, "Workshop on Foundations for Modeling and Simulation (M&S) Verification and Validation (V&V) in the 21st Century", better known as **Foundations '02**, was held October 22-24, 2002 in the Kossiakoff Conference and Education Center at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland (USA).

The process of validation was formally defined by the Department of Defense (DoD) and slightly modified by the American Institute of Aeronautics and Astronautics (AIAA) as

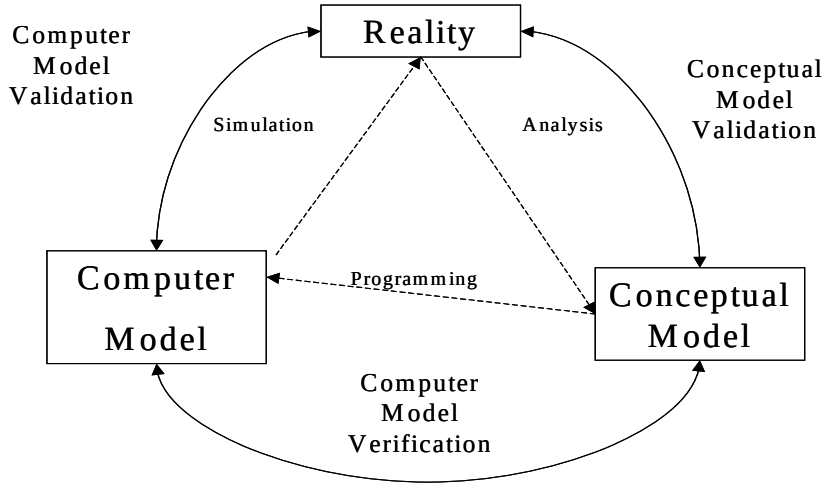
> *The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.*

An alternate term, *model assessment*, was proposed in (Fuentes et al (2003)). The latter term avoids the implication that the validation or assessment is done just once, to essentially certify a model's validity. Rather, model assessment is seen as part of an iterative process to continually improve models, based on an understanding of their capabilities and limitations at emulating reality. An important distinction to make is

[1] This white paper was developed in the summer of 2003 while serving a policy internship at the National Academies.

between the process of validation and the process of verification. Verification was defined by the DoD as *the process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model.* As illustrated in the diagram below (Cafeo and Cavendish, 2001), while validation is the process of comparing the computer model and conceptual model to reality, verification is the process of checking and debugging the code to make sure it reflects the conceptual model accurately.

Diagram from (Cafeo and Cavendish, 2001)



The goal of this paper is to give an overview of statistical methods and methodologies used or proposed for validation of complex computer models in different disciplines in sciences or social sciences. Examples of validation in transportation planning (Sacks *et al* 2000) and (Berk *et al* 2002), engineering, atmospheric sciences and social sciences will be used to illustrate the goal of validation and the methodologies.

In Section 1, computer models will be categorized in different types. General ideas about validation will be discussed in Section 2. In Section 3 statistical validation methods are discussed and a mathematical framework is introduced. Finally, some examples will be presented in Section 4.

1. Computer model types

The computer models considered in this paper are based on conceptual models that approximate reality and which are in turn represented by mathematical equations. For example, the spot welding example described in Section 4 (Bayarri et al (2002)) is based on a physical theoretical model which combines thermal, electrical and mechanical physics, and the CORridor SIMulation (CORSIM) (Sacks et al (2000)) microsimulation model is based on a stochastic model of traffic flow.

In a keynote address at the "Workshop on Statistical Approaches for the Evaluation of Complex Computer Models" (Berk *et al* 2002), Dr. William Press, the Deputy Director for Science and Technology at Los Alamos National Laboratories, proposed a taxonomy

and examples of computer models. Dr Press classified computer models in several types depending on the following aspects:

1. The conceptual model on which the computer model is based. The conceptual model is called "accurate" when the physics of the phenomena is known and deterministic, "statistically accurate" when the phenomenon is statistical, or "phenomenologically accurate" when the model captures qualitatively identifiable phenomena.
2. Type of input for the computer model. It is "accurate" when the input is a fixed value, "statistically accurate" when the input is a random variable.
3. Type of phenomena to be modeled. It is deterministic physical or emergent physical. An emergent physical phenomenon is neither explicitly represented in the system's elementary components or their couplings nor in the system's initial and boundary conditions.

The validation process will depend on the type of computer model. Dr. Press gave examples and comments on validation for some model types as summarized in the table below.

Computer model types	Examples	Comments on Validation
1. "Accurate" models of deterministic physical phenomena with "accurate" input conditions. 2. "Accurate" models of deterministic physical phenomena with "statistically accurate" input conditions.	Static civil engineering, models of bridges and dams, weapons code	- Three sources of modeling errors: error from conceptual model, error in the computer model (truncation vs round-off error), error from randomness of input for type 2. - Compare model run to data using appropriate norm.
3. "Statistically accurate" models of nondeterministic physical phenomena.	Turbulent fluid phenomena and climate models	- Need for better simulation methods to take into account uncertainty in the model. - Metric for model-to model and data-model evaluation
4. "Accurate" or "statistically accurate" models of emergent physical phenomena.	Statistical mechanics, smooth particle hydrodynamics and traffic flow modeling	
5. "Phenomenologically accurate" models of emergent physical phenomena. 6. "Phenomenologically interesting" models.	Turbulent intermittency, traffic jams and epidemics.	- Computer model generally inaccurate but still useful, could be used for training. - How to map "fields of data" into "phenomena" or "events" and the behavior of these phenomena.
7. "Video games" as models.		

A formal mathematical definition of input, output and parameters of a model is necessary to define the statistical validation framework, and this will be introduced in Section 3.2.

2. Validation

In addition to verification of the model code and checking that outputs of the model are considered reasonable by specialists in the field, validation of the model with real data is a necessary step for checking the accuracy. To validate a computer model, i.e., determine the degree to which the computer model is an accurate representation of the real world, results of computer model experiments need to be compared to real data. The comparison could be between the output of the model and past data, as for example evaluating a weather simulation model by comparing its output to past weather data (Covey et al. (2003)) or evaluating a microsimulation model for cancer treatment by comparing its output for patients with certain profiles to actual medical survey data for similar cohorts. The comparison could also be between the output of the model and the realizations of designed experiments or surveys, as for example in evaluating the predictions from car crash models by comparing them to the outcomes of designed car crash experiments. Such validation by data is called external validation in the social sciences.

Although several replicates are necessary to account for the variability[2] in the data from the phenomena of interest, the cost of collecting and processing the data often limits the number of available replicates of the phenomena. For example, collecting and processing data to validate CORSIM required manual recording of traffic and video coverage, and it was very costly. Performing real car crash experiments with dummies in order to validate a car crash model is costly, and collecting new data or retrieving archived data sets from past history to validate a social or economic model is also costly.

When an appropriate set of data is collected, comparison with the computer model will involve analyzing the error or bias, i.e., the difference between the model prediction or forecast and the field data. The variability in the data and the randomness and sources of error in the model are not always well understood. As discussed in the internal research report of General Motors (Cafeo and Cavendish (2001)):

> *A major problem with the use of math models and simulations in support of product and process design is that the models are only the abstractions of reality, and the insights and understanding they can provide is limited. It is important that the model builders and code users understand the limitations of these models used to support product and process design. It is unfortunate however, that almost all computational models used to support engineering design are used deterministically, that is they are seldom exercised to explicitly account for error and uncertainty, and they do not provide boundaries on the range of valid model applications.*

[2] Sources of variability of the data are measurement or sampling errors and inherent variability of the phenomena under different conditions.

The error in the model due to uncertainty in the input or parameters in the model is often not accounted for because some input or parameters are considered fixed, as discussed in Section 4.2.3 of (Cafeo and Cavendish (2001)):

> *The standard engineering practice is to estimate, one way or another (literature, mean values of data obtained from experiments, etc), a single value for such parameters and proceed with the calculation using these "representative" values. This may be an adequate approach to treating this uncertainty, especially if it can be argued that the range of uncertain parameters values is narrow and computed results are not sensitive to variations in these uncertain parameters. When parameter uncertainty is important, then we argue that the calculations made with best estimates of single values of uncertain parameters are not the appropriate way of dealing with this uncertainty – especially when making comparisons of computed results with data derived from validation experiments.*

Considering that the input and parameters are fixed ignores the uncertainties of these values and thus the propagation of these uncertainties to the model output that accounts for the variability of the error. Understanding the variability of the error would allow a proper measure of how close the output of a model is to the real data in a statement like: "Based on the analysis of these validation experiments and comparisons with computations, we are 80% confident that the actual system or process performance will differ from the computational prediction by no more than 10%." (Cafeo and Cavendish (2001))

To obtain an accurate description of the variability or distribution of the error often necessitates multiple runs of the model. Time and the corresponding cost, for running multiple simulations of a computer model depends on how many inputs and how complex the model is. So, even when field data is abundant—as is the case for the atmospheric sciences, where some atmospheric data has been collected hourly or daily for many years and for a variety of locations—the complexity of the model may still make validation problematic, as noted in (Fuentes et al (2003)):

> *Evaluation of the performance of a numerical model is mostly constrained by the amount and quality of observational data available for comparison with modeling results, and by the ease with which the models can provide runs that are appropriate to compare to the data.*

Validation with scarce data, simplification of complex models for simulation of the error, distribution of the error and decomposition of the error are discussed in the steps of the validation framework in Section 3.1 A mathematical framework to address the uncertainties and define the error will be presented in Section 3.2.

3. Statistical Validation

3.1. Validation Framework

Validation steps depend on the model that is being validated and the goal of the

analyst. However, one general framework for validation has been proposed in several papers. In particular, the technical report (Bayarri et al (2002)) by NISS in collaboration with General Motors proposed a six-step iterative framework for validation and applied it to two test beds: a car crash model and a spot welding model. The latter will be presented in Section 4. The six steps are described in this Section with a few comments and recommendations from other papers.

The six-step validation procedure in (Bayarri et al (2002) and (Cafeo and Cavendish (2001)) is an iterative one:

> *…a series of activities or steps. These are roughly ordered by sequence in which they are performed. The completion of some or all in the series of activities will typically lead to new issues and questions, requiring revision and revisiting of some or all the activities, even if the model is unchanged. New demands placed on the model and changes in the mode through new development make validation a continuing process. The framework must allow for such dynamics. (Quoted from Cafeo and Cavendish (2001).)*

The six steps are:

1. Specifying model inputs and parameters with associated uncertainties or ranges---the Input/ Uncertainty (I/U) map.
2. Determining the evaluation criteria.
3. Collecting data and designing experiments.
4. Approximating output of the computer model.
5. Analyzing model output; comparing computer model output with field data.
6. Feedback information into current validation exercise and feed-forward information into future validation activities.

Step 1

In the first step, an assessment of uncertainties in the model inputs and parameters (e.g., fixed vs. variable, known vs. unknown, and range of uncertainty) is done by experts. When the number of inputs is large, it is essential to set priorities among the inputs in this step to help design the experiments, survey or collection of data.

Step 2

In the second step a choice is made for specific evaluation criteria to compare the model to reality. These criteria could be particular outputs, all outputs or a function of the outputs. The data collection will be also affected by this step and, as the data are analyzed, the evaluation criterion might be revisited in the iterative process.

Step 3

The first and second steps allow the design of informative experiments or data collection in the third step. The data collection can be done in multiple stages; at each stage different scenarios varying one input at a time or a block of inputs or modules of the model are considered. When the number of inputs is large or some input's variation range

is large, several papers suggested using "space-filling" strategies of choosing the input values at which to experiment. One such method is the Latin Hypercube Design[3]. The end of each data collection stage is the end of an iteration of steps 1-3, i.e., after each stage, the steps1-3 would be reconsidered before the next stage.

Step 4

While other steps are necessary in validation, this step is optional. Approximating the computer model by a faster model in this step would save time in simulations when the original model is not fast enough (G. Molina et al. (2003)) and (Fuentes et al. (2003). A few statistical techniques are proposed in (Bayarri et al. (2002)) to approximate a model including dimensionality reduction techniques which identify and exclude less significant elements in the model (e.g., Principal Component methods and Proper Orthogonal Decomposition such as 'ANalysis Of VAriance' (ANOVA)); linearization/Gaussian error accumulation method, which linearizes the model so that input distributions can be passed though the model using linear Gaussian updating; response surface methodology, including Gaussian processes (used in (Bayarri et al. (2002)) and neural networks; and Bayesian networks, which allow uncertainty transference between sub-models from which the model is constructed. An approximating model is called a *meta-model* or an *emulator* in (Fuentes et al. (2003)), and methods for generating an emulator of atmospheric sciences models are discussed in this paper.

Step 5

The outcome of the validation is determined in the fifth step wherein the comparison of model output to reality takes place. One first performs a sensitivity analysis on the model, which does not rely on real data. The goal of sensitivity analysis is to understand the propagation of the uncertainties from the model input to the model output and to determine which inputs affect the output more strongly. Then the model output is compared to the output from field data, the error is decomposed into multiple sources of error (e.g., random measurement error in data collection, error in tuning or/and calibration[4], error in the model's description of reality). Visual tools such as the graphics used in (Covey et al. (2003)) or the animation of CORSIM could help determine the sources of error and visualize the mean and the variation of the error.

Step 6

This last step is the iterative step. Information from previous steps is used to improve the model and the improved model is subsequently validated through steps1-5.

---

[3] For independent inputs, the idea of the Latin Hypercube design derives from the field of Latin square experimental deign. For a discussion of this method, refer to McKay, M. Beckman, R., and Conover, W., (1979), "A comparison of Three Methods of Selecting Values of Input Variables in the Analysis of Output from a Computer Code", *Technometrics*, Vol.21, #2. 239-245.

[4] From (Sacks et al. 2000) Calibration and tuning a model are general terms, often used interchangeably. Tuning is a phrase commonly associated with adjusting input parameters to match model output whereas calibrating refers to the process where the model output are used, either alone or with field data, to determine input parameters. "In calibration, one tries to find the true---but unknown---physical value of the parameter, while in tuning one simply tried to find the best fitting value"

3.2 Mathematical Framework

To specify the error and separate well-known fixed or variable input of the model from calibrated or tuned input a useful framework was used in (Bayarri *et al.* (2002)), (Trucano *et al.* (2001)), (Easterling and Berger (2002)), and (Fuentes et al (2003)). The error is defined as the arithmetic difference between the output or numerical result of the model and the corresponding output from reality (past data, experimental data or survey): "error = model - reality". "Inputs" denoted by $\boldsymbol{x}$ are distinguished from "parameters" denoted by $\boldsymbol{u}$, where $\boldsymbol{x}$ and $\boldsymbol{u}$ are necessary for the model $\boldsymbol{M}$ to compute the output $\boldsymbol{y_M}$. The output from reality corresponding to the same input $\boldsymbol{x}$ is denoted by $\boldsymbol{y}$,

$$\boldsymbol{y_M = M(x\,;\,u)}$$
$$\boldsymbol{y = y_M + e(x)},$$

where $\boldsymbol{e(x)}$ is the unknown error or bias of the model and $\boldsymbol{x}$ is the vector of controllable inputs (Bayarri *et al.* 2002), a function of space and/or time (Trucano *et al.*, 2001), or the set of variables whose values define a physical entity and the environment to which it is subjected. For example, $\boldsymbol{x}$ might represent physical dimension(s), materials, environment variables, and/or initial boundary conditions (Easterling and Berger, 2002). On the other hand, the model parameter $\boldsymbol{u}$ is the vector of unknown tuning and/or calibration parameters in the model (Bayarri *et al.* 2002). It includes parameters that are needed to specify physical responses in the models, such as transfer coefficients in the set of equations on which $\boldsymbol{M}$ is based (Easterling and Berger, 2002).

The error $\boldsymbol{e(x)}$ contains errors from the uncertainty of the input in the model and possible model error. Note that because $\boldsymbol{y}$ can't always be known exactly due to measurement errors or mismatches between physical testing and model structure, it is often difficult to characterize $\boldsymbol{e(x)}$. Investigating $\boldsymbol{e(x)}$ over *ranges* of x of interest, for example by looking at the distribution of $\boldsymbol{e(x)}$ and its mean and variance, would allow evaluating model accuracy and the model's predictive capability. The error is a function of the inputs. If the error is a linear function of the inputs, then subtracting the linear regression of the error on the inputs from the model would correct for the bias. This method is often used in engineering when the existence of a bias is known. Subtracting a nonlinear function of the input from the model to account for the bias was done in (Bayarri et al. (2002)).

4. Examples

Several examples from engineering, atmospheric sciences, and social sciences are cited in this paper. Three validation examples will be described in more detail and others will be described very briefly. The first example is a resistance spot welding model (Bayarri et al. (2002)) which was validated using the six step procedure defined in Section 3.1. The second example presents the validation of CORSIM, a microsimulator of traffic flow in a street network. The third example is the comparison and evaluation of 18 models in atmospheric sciences. Other examples described briefly are from engineering, atmospheric sciences, and health sciences.

Resistance spot welding model: The physical theoretical model of spot welding combines thermal, electrical and mechanical physics. It is a coupling of partial differential

equations that govern heat and electrical conduction with those that govern temperature dependant, elastic/plastic mechanical deformation. The inputs include the geometry, material properties, conductivities, electrical resistivity, numerical parameters, current and load. The output of interest is the diameter of the resulting weld nugget.

1. I/U map: The I/U map displayed in a table informs that the first three inputs (geometry, material properties and conductivities) are varied, the electrical resistivity is a tuned parameter, the numerical parameters are set to default values and the current and load are fixed.
2. Evaluation criteria: the two outputs of interest represent the evaluation criteria: nugget size after 8-cycles and nugget size as a function of the number of cycles.
3. Data collection and design of experiments: Because there are many inputs, some fixed and some variable, and the variable inputs are either discrete or continuous, it is impossible to test for all possible values of the input. Therefore, the Latin Hypercube Design was used to design 35 different experiments.
4. Approximation of computer model output: To approximate the model by a random function, a Gaussian process response surface approximation (GASP) was used. In order to use the same field data for tuning the parameter and validating the model, a Bayesian GASP was used.
5. Analyses of model output; comparing computer model output with field data and Feedback loop: using the Bayesian formulation, the bias of the model was estimated along with the distribution of the tuning parameter and uncertainty tolerances on the bias function and predictions were calculated.

Conclusion of the validation: The posterior distribution of the resistivity parameter shows a high uncertainty. The model has a bias, and the bias remains even after tuning. However, the bias-corrected predictions might be tolerable.

Transportation example

The papers by (Sacks et al. (2002)), (Sacks *et al.* (2000)), and summarized presentation of Nagui Rouphail, Jerome Sacks and Byungky Park in (Berk *et al* (2002)) present the CORridor SIMulation (CORSIM) and its statistical validation. CORSIM is a microsimulation computer model that simulates traffic flow in a street network under complex conditions, including traffic signal settings. The two main questions which motivated the validation are: how well does CORSIM reproduce field condition and how well does CORSIM predict new situations?

In addition to these papers, a working paper by (G. Molina et al (2003)) presents a method for Bayesian tuning of CORSIM.

To address the validation questions, the National Institute of Statistical Sciences (NISS) undertook a case study with the cooperation of the Chicago Department of Transportation and the Urban Transportation Center of the University of Illinois at Chicago. Data were collected on an important street network in the city of Chicago. This data was used both for determining the values of some inputs to CORSIM and also to evaluate CORSIM's capability to model field conditions.

CORSIM is a stochastic simulator that moves vehicles second-by-second through a network. It represents individual vehicles (hence the name microsimulator) which enter the road network at random times, move according to local interaction rules describing governing phenomena, such as vehicle following and lane changing, and turn (or not) at intersections according to prescribed probabilities.

1. I/U map: Inputs are classified in three types: fixed and controllable inputs, random and noncontrolllable inputs and controllable inputs. The fixed and controllable inputs include the geometry (link and node) of the street network (e.g., distance between intersections, number of traffic lanes), the placement of stop signs, bus stops and routes and parking conditions. Random and noncontrollable inputs include generation of vehicles by sampling inter-arrival time distributions at each entry node (parameters for the inter-arrival times were estimated by a simple moment estimator of a parameter of a gamma distribution) and designation of vehicle type (auto or truck) by making independent Bernoulli trials with a fixed probability estimated from field data. The dwelling time of buses at bus stops and inter-arrival times at entry nodes are also considered random. Other random parameters are turn probabilities (estimated from field data) and driver characteristics such as car-following behavior and lane-changing maneuvers, for which CORSIM provides default distributions. Finally, controllable inputs include settings of the traffic signals, such as cycle length, green times and offsets.
2. Evaluation criteria: CORSIM provides several outputs: an animation package that enables the visualization of the traffic movements and aggregated numerical output for each link. The latter includes the number of trips on each link, average link travel time, link queue time, maximum queue length on each lane in the link, and link delays. The evaluation relies on an evaluation function and comparison of animation output against real video. The evaluation function used in the three papers (Sacks et al. (2000)), (Berk et al. (2002)) and (G. Molina (2003)) differ. In the paper by (Sacks et al. (2000), the stop-time on approaches to intersections was used as the primary evaluation function. In (Berk et al. (2002)), the maximum queue length (MQL) was used as an evaluation function. Finally, in (G. Molina (2003)) the total queuing time of vehicles in the network was used. A visual validation by using the animation was used to check assumptions in the model.
3. Data Collection: Data was collected on an important street network in the city of Chicago. The data collection was either through observers or video recording. The data was processed for three time periods of an hour each covering “peak” as well as “shoulder” period. This data was used for tuning parameters in CORSIM and also to evaluate CORSIM’s capability to model field conditions.
4. Approximation of computer model output: Because the computer model is not fast enough to apply the Bayesian tuning using Markov Chain Monte Carlo approach, a simpler stochastic network that mimics the traffic simulator with respect to the two tuning parameters of interest was proposed in (G. Molina et al. (2003)).
5. Analysis of output and Feedback: When high variability was found in the evaluations function, the simulations were further explored, which led to a better tuning of the parameters to reflect the conditions in the field. For example, the

histogram displaying the distribution of MQL for 100 simulations in (Berk et al. (2002)) shows that the field data MQL is in the range of variation. However, the variability of MQL in the simulations is high and causes spillback and gridlock[5] not observed in the field. To check the reason for such high variability, the animation was consulted and the cause was determined to be long stopping time at a stop sign in the model compared to the "rolling-stop" behavior in the field data. The model was adjusted for a lower stopping time to account for this behavior and the variability of the mean queue time was significantly reduced, resulting in an absence of spillback. Similarly, the speed was changed from 30 to 20 miles per hour in (Sacks et al. (2000)) to be more consistent with the field data.

Conclusion: CORSIM is imperfect but can be used effectively to plan signals in an urban road network.

Atmospheric science example (Curt Covey *et al.* (2003))
This report presents a comparison and evaluation by the Coupled Model Intercomparison Project (CMIP) of 18 atmospheric models developed by different research groups. CMIP was established in 1995 in an effort to understand why some atmospheric models using global coupled ocean-atmosphere general circulation models (coupled GCMs)[6] developed by different research groups were giving somewhat contradictory answers to the same questions. In particular, models were giving different answers to the questions involving the effect of increase of CO_2 on global warming. The differences in the output come partly from different assumptions and adjustments that the models make.

In (Covey et al. (2003)), the simulations of the different models are compared to each other and also compared to the measured values over an 80-year period up to the present. Results of simulations from these 18 models and variations of the simulations and errors were presented using several visualization tools. More specifically, the paper used time series plots, latitude-longitude and latitude-height plots with mean contours and shaded variance, a Taylor diagram, and space-time error plots.

The inputs to GCMs include a small number of external boundary conditions such as the solar "constant" and atmospheric concentration of radiatively active gases and aerosols. The outputs analyzed in this paper include surface air temperature, precipitation, mean sea level pressure, humidity, ocean temperature at 1000 m depth, barotropic stream function, and sea ice thickness. These outputs vary over time and space.

Some of the analysis and corresponding graphics are described below

1. Comparison of global- and annual mean observations. Observations and results of simulations are averaged in both time (average of monthly means to form an annual mean) and location (average over latitude and longitude of the models

[5] "Spillback occurs when congestion causes traffic to back up and block movement at an upstream intersection. Failure of spillback to clear up can result in gridlock."
[6] GCMs that include interactive sea ice simulate the physical climate system, given only a small number of external boundary conditions such as the solar "constant" and atmospheric concentration of radiatively active gases and aerosols.

simulations). Two time series displayed the global- and annual temperature means over the last 80 years and the global- and annual precipitation means for the last 80 years. “The most striking aspect is the stability of model-simulated temperature and precipitation.”

2. Comparison of long-term mean averages (average for the past 80 years). The long-term time mean averages for most inputs are presented in a latitude-longitude axis plot or/and in a latitude-and-height axis plot to visualize the zonal variations. For each choice of axis, there are four panels. One panel allows visualization of the variations in the model simulation by representing the average over all models of the long-term mean average (contour) and the intermodel standard deviation (color shading). A second panel allows visualization of the variations in the error, i.e., the difference between a model’s mean and the observations, by presenting the mean error (contour) and the intermodel standard deviation of the error (color shading). A third panel gives zonal averages for the individual model control runs and the observations, each represented as a curve. To control for the variation over time, a last panel gives the average over all models of the difference between the last 20-year mean and the first 20-year mean from the 80-year perturbation simulations, in which atmospheric carbon dioxide increases at a rate of 1% per year (contours), together with this difference normalized by the corresponding standard deviation (color shading).
3. Taylor diagrams of the total spatial and temporal variability of three fields. A Taylor diagram was constructed for surface air temperature, sea level pressure and precipitation. This diagram allows the visualization of three quantities—standard deviation normalized by observation, correlation with observation, and root mean square difference from observation—in a two-dimensional space7.
4. The component of space-time errors diagram: it is a shaded table, where each column is a model and each row is a component of the error. The total error of each model was decomposed into bias and pattern error. The bias is the error component associated with global- and annual mean and the pattern error is the remaining error. Instead of numbers, the table displays a color coded scheme varying from blue (lowest) to red (highest) value of the error component.
5. Other graphics include the spectral density8 by years for observed and models simulations as well as the confidence limit.

The main result of this analysis is that all models “simulate an overall level of natural internal climate variability that is within the bounds set by observations.”
In this atmospheric science example, the validation data are not experimental, but they are the available atmospheric history. Hence, this study doesn’t allow to track meaningful

[7] “The radial coordinate is the ratio of the modeled to observed standard deviation. The cosine of the angle of the model point from the horizontal axis is the spatio-temporal correlation between model and observation. When plotted in these coordinates, the diagram also indicates the root-mean-square difference between model and observation: this difference is proportional to the linear distance between the model point and the ’observed’ point lying on the horizontal axis at unit distance from the origin.”

[8] In this paper, “The spectral density follows the algorithms described by Jenkins and Watts calculating the spectra from autocovariance with lags up to 1/4 the length of each time series and using a Tukey window 1/10 the length of each time series.”

structural changes in the system (here he environment, e.g. due to different carbon emissions) which we often hope for in validation.

Other examples

In the SANDIA validation metric projects paper by (Trucano et al. (2001)), methods for validation of two examples were discussed: a structural dynamic case study and foam decomposition case study. The first case study is related to weapons simulations in normal Stockpile-to-Target-Sequences (STS) environments, and the second case study considers the problem of foam decomposition under thermal states related to fire relevant to abnormal STS environments.

In (Fuentes et al. (2003)), instead of using the term validation, authors used the term "model assessment" to describe the same procedure. This paper presents several approaches for applying statistical techniques to model assessment, applies the approaches to atmospheric models, and compares the approaches. The Bayesian melding technique is illustrated when there is ample model output and sparse monitoring data by applying it to the atmospheric model called "Model-3". The geostatistical approach is illustrated when monitoring data is required from a dense network but detailed analysis of the model output is enabled by applying it to the SARMAP atmospheric model. The last approach describes a situation where it is difficult to get a sufficient number of model runs and a statistical approximation to the model output is compared to data. This last approach is applied to the SACCO atmospheric model.

Other useful discussions of statistical validation of computer models appear in (Davis et al. (2000)), which presents a validation of the Regional Oxidant Model (ROM); in (Berk et al. (2002)), which presents methods for validation in meteorology, wildfire control and immune system function (in addition to the CORSIM discussion already mentioned); and in two working papers from Statistics Canada (Flanagan et al. (2003) and Edward Ng et al. (2002)), which discuss the validation of the POpulation HEalth Model (POHEM) microsimulation model for cancer screening. In Addition, a car crashed model was validated using the six-step validation procedure in (Bayarri et al (2002)).

References

1. Foundations'02 V & V Workshop, https://www.dmso.mil/public/transition/vva/foundations
    a. Robert G. Easterling, James O. Berger (2002) "Statistical Foundations of the Validation of Computer Models".
    b. William L.Oberkampf, Timothy G. Trucano, Charles Hirsch (2002) "Verification, Validation, and Predictive Capability in Computational Engineering and Physics".

2. M. J. Bayarri, James O.Berger, David Higdon, Marc C. Kennedy, A. Kottas, Rui Paulo, Jerome Sacks, James A. Cafeo, James C. Cavendish, C.H. Lin, and J. Tui (2002), "A Framework for Validation of Computer Models". Technical report number 128, NISS. http://www.niss.org/technicalreports.html

3. Richard A. Berk, Peter Bickel, Katherine Campbell, Robert Fovell, Sallie Keller-McNulty, Elizabeth Kelly, Rodman Linn, Byungkyu Park, Alan Perelson, Nagui M. Rouphail, Jerome Sacks and Frederick Schoenberg (2002)," Workshop on Statistical Approaches for the Evaluation of Complex Computer Models", *Statistical Sciences*, Vol. 17, No. 2, 173-192.

4. John A. Cafeo, James C. Cavendish (July, 2001), "A Framework for Verification and Validation of Computer Models and Simulations", Internal Research Report at General Motors research and development center.

5. Curt Covey, Krisha M. AchutaRao, Ulrich Cubasch, Phil Jones, Steven J. Lambert, Michael E. Mann, Thomas J. Phillips and Karl E. Taylor (2003) "An Overview of Results from the Coupled Model Intercomparison Project". *Global and Planetary Change*, 37(2003) 103-133.

6. Davis, J. M., Nychka, D. and Bailey, B. (2000). "A Comparison of the Regional Oxidant Model with Observed Data". *Atmospheric Environment*, 34, 2413-2423.

7. William Flanagan, Christel Le Petit, Jean-Marie Berthelot, Kathleen J. White, B. Ann Coombs, Elaine Jones-McLean (2003). "Potential Impact of Population-based Colorectal Cancer Screening in Canada." Working paper, Statistics Canada. http://www.statcan.ca/start.html

8. Montserrat Fuentes, Peter Guttorp and Peter Challenor, "Statistical Assessment of Numerical Models" (2003), Technical report, National Research Center for Statistics and the Environment (NRCSE), University of Washington.

9. M. Granger Morgan and Max Henrion (1990), *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, Cambridge, UK.

10. G. Molina, M.J Bayarri, J.O. Berger (2003) “Statistical Inverse Analysis for a Network Microsimulator” Working paper, Institute of Statistics and Decision Sciences (ISDS) Duke University.

11. NAS (1991) *Improving Information for Social Policy Decisions -- The Uses of Microsimulation Modeling: Volume I, Review and Recommendations*, report of the Committee on National Statistics, National Research Council, Washington DC.

12. NAS (1998) *Statistics, Testing, and Defense Acquisition: New approaches and Methodological Improvements*, report of the Committee on National Statistics, National Research Council, Washington DC.

13. Jerome Sacks, Nagui M. Rouphail, B. Brian Park, Piyushimita Thakuriah (2002), “Statistically-Based Validation of Computer Simulation Models in Traffic Operations and Management”. *Journal of Transportation and Statistics*, Vol. 5, pp.1-15.

14. Jerome Sacks, Nagui M. Rouphail, B. Brian Park, Piyushimita Thakuriah (2000), “Statistically-Based Validation of Computer Simulation Models in Traffic Operations and Management”. Technical report, NISS 112. http://www.niss.org/technicalreports.html

15. Edward Ng, William Flanagan, Jean-Marie Berthelot, Simone Dahrouge, Jean Maroun (2002)“Survival Parameter Estimation and Validation on a Colorectal Cancer Disease Progression Model”, working paper, Statistics Canada. http://www.statcan.ca/start.html

16. Timothy G. Trucano, Robert G. Easterling, Kevin J. Dowding, Thomas L. Paez, Angel Urbina, Vicente J. Romero, Brian M. Rutherford, and Richard G. Hills (2001), “Description of the Sandia Validation Metrics Project”, Sandia National Laboratories, SAND2001-0243