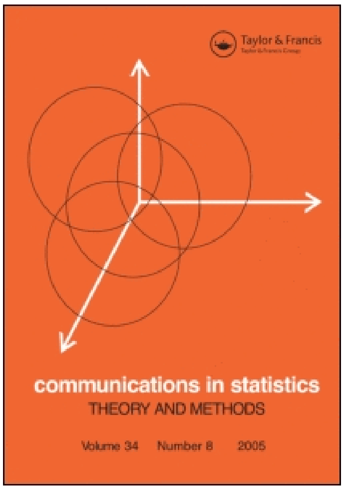

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

Model validation: an annotated bibliography

Martin A. Hamilton[a]
[a] Department of Mathematical Sciences, Montana State University, Bozeman, MT, U.S.A.

To cite this Article Hamilton, Martin A.(1991) 'Model validation: an annotated bibliography', Communications in Statistics - Theory and Methods, 20: 7, 2207 — 2266
To link to this Article: DOI: 10.1080/03610929108830628
URL: http://dx.doi.org/10.1080/03610929108830628

COMMUN. STATIST.-THEORY METH., 20(7), 2207-2266 (1991)

MODEL VALIDATION: AN ANNOTATED BIBLIOGRAPHY

Martin A. Hamilton

Department of Mathematical Sciences
Montana State University
Bozeman, MT 59717-0240
U.S.A.

Key Words and Phrases: mathematical model; computer model; validation; evaluation; assessment; analysis; verification

ABSTRACT

This paper contains an annotated bibliography of 316 papers and books relevant to model validation, with emphasis on articles of potential interest to statisticians. The bibliography is accompanied by a brief introduction, a list of some important areas for statistical research, a short glossary, and a key word index.

1. INTRODUCTION

"The map is not the territory" (Einhorn and Hogarth, 1982).

1.1 Mathematical and Computer Models

Mathematical and computer modeling has become an indispensable tool for policy makers and scientists. Policy decisions on important issues, such as the greenhouse effect and the disposal of radioactive waste, have come to rely heavily on computer simulations. Most scientific advances depend on models – models as expressions of theory, models that provide a way for the theory to be compared to data, and models that lead to new discoveries. The list of modeling successes is long, and will rapidly become longer as specialists develop more sophisticated quantitative methods and have access to faster, cheaper computer equipment. Pool (1989) has suggested that computer modeling/experimentation "will eventually grow into a third domain of science, coequal with the traditional domains of theory and experimentation."

1.2 Model Validation

There is no clear consensus on terminology in the modeling literature, and the words "validation", "verification", "evaluation", etc., mean different things to different people. For purposes of this paper, **"model validation"** is an assessment of the extent to which a model is well-founded and fulfills the purpose for which it is formulated (Cobelli et al. 1984). Many experts prefer not to use the term validation because of its connotation of truth (e.g., Swartzman and Kaluzny 1987). Among other considerations, model validation requires three important tasks, which I shall call verification, sensitivity analysis, and evaluation.

"Verification," as used here, is determination of the extent to which the numerical model or computer model is a faithful representation of what was intended in the mathematical model and that the mathematical model is a faithful representation of the underlying theory. The fields of software engineering and numerical analysis are often involved in verification activities.

"Sensitivity analysis" is determination of the extent to which model behavior (or behavior of each component of the model) is affected by changes in parameters. Often the parameters are determined by statistical estimation at the calibration phase of model building. Statistical estimation techniques, error propagation analyses, and Monte Carlo experimentation provide important contributions to sensitivity analyses.

"Evaluation" is a comparison of the model's output to measurements on the real system. Statistical theories of estimation, significance testing, and ranking/selection provide important evaluation tools.

Because models are so influential, it is prudent to require that each model is submitted to careful examination before it is given credibility. Model predictions or forecasts should always be accompanied by a measure of uncertainty to indicate model validity as well as any inherent, stochastic error. No model exactly represents the real world, and users must interpret a model's prediction in light of the uncertainties associated with that prediction. As we gain knowledge about the true mechanisms governing the natural system, the structure of the present model will be replaced by a better one. Meanwhile, it is crucial to have some knowledge of the model's relationship to reality. Nevertheless, there seems to be a natural tendency for people to accept information produced by a computer, even if the reliability of the underlying model has not been determined. The statistician G.E.P. Box recently commented that "students absorb the impression that models are true and data are wrong instead of the other way around."

Time and resources are required for model validation. But there are pressures on model builders to skimp on validation activities. As Hodges (1987) wrote, "The analysis is usually done under time and budget pressure...not to produce any particular result, but to produce some result, and analytical niceties are expected to yield. This makes it difficult to get funding for model criticism and improvement, and fosters an atmosphere in which scrupulous model criticism is viewed as vaguely traitorous."

Because measurement of a model's validity is dependent on the purpose for the model, model users should work closely with the model builders to decide on model validation specifications. Then model validation considerations can be integrated into the model building process from start to finish. There is a close analogy to quality assurance programs for manufactured products. Those involved in model validation might well benefit from knowledge of the principles of quality assurance.

In some instances, as when the model exists in the form of a large, complicated computer program or when someone is considering a use different from that for which the model was built, the task of conducting a complete validation may be assigned to a third party, who is neither the model builder nor the model user. Third party validation is called assessment. Model assessors (also called model analysts) are concerned with model validation in the abstract as well as with the particular model at hand. They contribute by studying model validation procedures, by giving potential users a place to turn for impartial analyses, and by promoting a generally higher level of performance in the modeling trade (Richels 1981; chapter by Greenberger in Gass 1980).

Because modeling is used in many fields and models are created for a wide variety of purposes, the literature on model validation is quite diverse. Although the bulk of that literature is narrowly focused on individual models, there are many articles of general interest to model validation analysts. Few of those articles are in the statistical literature. On the other hand, there are a number of articles in the statistical literature which introduce statistical methodology that is directly applicable to model validation work, even though the authors may not necessarily have model validation in mind. It seems useful to gather together those citations on model validation that are of potential use to the community of statisticians.

1.3 Goal of this Paper

The main goal of this paper is to facilitate access to the literature on model validation by providing an extensive bibliography in which each citation contains a brief annotation and a relevant set of key words. Published articles and technical reports containing information on the evaluation of mathematical/computer models number in the thousands. Most of those articles, however, are concerned with the evaluation or verification of a specific model, not with general principles. This bibliography focuses on methodology papers as opposed to papers on a specific model, although some selected specific-model papers are cited. Without doubt, some relevant citations have been overlooked, but I am confident that the major components of model validation are identified and discussed in the collection of books and articles presented here. For most topics, only the most recent articles have been included in the bibliography.

To identify the most useful articles published in English since 1980, I conducted searches in major on-line bibliographic data bases. In addition, experts in various fields of science, mathematics, statistics, and engineering were asked to recommend papers on model validation. Then, starting with articles so identified, I followed the literature associated with the most instructive articles. I ceased entering citations to the bibliography in July, 1990. The bibliography contains less than half the papers I read for possible inclusion. It references 93 different journals and 500+ authors. The key words and annotations in the bibliography do not necessarily emphasize the main points of the associated paper; instead, they focus on the aspects of the paper related to model validation. Consequently, the main scientific contributions of the paper may be quite different from the key words and annotation.

To help the reader use the bibliography with its key words and annotations, Section 2 provides a review of the important components of model building and the associated terminology. Section 3 is a list of some aspects of model validation that would benefit from new statistical theory and methodology. Following the Bibliography are two appendices. Appendix A is a glossary and Appendix B is an index that links key words to citations.

2. COMPONENTS OF THE MODELING PROCESS

A theory of model building does not seem to exist, although the review of Costanza and Sklar (1985) suggests the beginnings of a theory. The primary structure underlying model validation is prescriptive; that is, it consists of a checklist of steps that should be performed. Many of the papers cited in the bibliography contain checklists and flow charts. To convey the flavor of this prescriptive approach, I have divided the modeling process into three basic components: input, structure, and output. A list of these components, and the main subcomponents, is presented in Table 1. A validation report would describe the validity and uncertainty of each of those items.

Because the terminology of model validation is not standard, it is possible that some of the key words in the bibliography and phrases in Table 1 may be subject to various interpretations. For this reason, Appendix A is a glossary of the terminology as I am using it.

3. STATISTICAL ISSUES DESERVING ATTENTION

A complete model validation requires more than verification, sensitivity analysis, and evaluation; e.g., judging face validity, checking the quality of data gathering and processing systems, determining the degree of access to, and usability of, the model and its documentation, and listing suggestions for model improvement.

When one considers all the potential errors and sources of uncertainty that might affect a mathematical/computer model, one finds that each branch of statistics can be applied to model validation. Thrall (1989) identifies model validation as one of the major challenges facing the statistical community, and he describes many potential statistical contributions to model validation activities. Statisticians can expect to play a prominent role. Among areas for which new statistical methodology would be especially helpful, I have identified ten priority areas; they are listed below with references to relevant papers.

1. **Lumping/aggregating** – determining effects of lumping or aggregation on evaluations and evaluation methodology [Beven 1989; Hayes and Moore 1986; Switzer 1989].

2. **Temporal/spatial correlations** – When evaluating model predictions against real system data, the statistical analysis must account for these potential correlations. At present, such correlations are too often ignored. [Beran 1989; Cliff and Ord 1981; Feldman et al. 1984; and articles possessing key words: "cross-correlation," "serial correlation," "spatial correlation," "spatial model," or "time series"]

3. **Extreme values** – methods for assessing the performance of the model when the model will be used operationally to predict extreme values such as the maximum, 99th percentile, etc. [Downton and Dennis 1985; articles possessing key words "extreme events."]

4. **Matching** – In evaluation, it is not always clear how to match predictions with observations; consider air pollution models, for example, where predictions can be made on a dense grid, but observations are available at only a few scattered sites. This is an especially thorny problem when interest is in extreme values. [Dennis and Downton 1984; Rosenbaum and Rubin 1985; Joglekar et al. 1989; Tesche et al. 1987; articles possessing key word "matching"].

5. **Ranking and selection** – In many circumstances, the model user wishes to choose the best from among several competing models. Ranking and selection techniques applied to goodness-of-fit statistics seem apropos, but they are seldom used in this context. [Bush and Mosteller 1988; Duông 1989; Hjorth 1989; Novotny and McDonald 1986; articles possessing key words "ranking and selection"].

6. **Graphical** – In many modeling papers, the sole evaluation is a visual comparison of predicted to observed. Modern graphical techniques should be designed and adapted for this purpose. [articles possessing key word "graphics"]

7. **Distribution-free statistical methodology** – There is a need for distribution-free approaches to all aspects of model validation. Bootstrapping has much to contribute here. [Azzalini et al. 1989; Deshpande & Mehta 1983; Dingman & Bedford 1986; Eubank & Spiegelman 1990; Lam 1981; Peters & Randles 1990; Raynor & Best 1990; articles possessing key words "bootstrapping" or "jackknife"].

8. **Expert opinion** – How to choose experts; how to elicit their opinions in a form suitable for modeling and model validation; how to determine the uncertainty of expert opinions. [articles possessing key words "expert opinion"].

9. **Uncertainty** – How to characterize uncertainty for model users, especially when overall uncertainty is dependent on the uncertainties of many individual modeling components. [articles possessing the key word "uncertainty"].

10. **Philosophical bases for model validation** – There are many general questions for which no consensus seems to exist. Consider these examples. (i) When, if ever, can a model be rejected or confirmed? (ii) How are model users supposed to apply model validation information? (iii) How can separate model assessments be pooled (meta-analysis)? (iv) What are the relative merits of graphics, statistical estimates, significance tests, etc. for model validation purposes? (v) When, if ever, should the analysis be frequentist, Bayesian, or empirical Bayesian? [Belsley 1988; Finkel 1990; Freedman 1985; articles possessing key words "philosophy," "general," "testing," and "Bayesian"].

These are some of the areas where statisticians can make contributions to the theory and methodology of model validation. At the application end of the process, it is important to train policy makers, scientists, journalists, and the public to view mathematical/computer model predictions with less credulity. Statisticians should explain the importance of model validation activities in the classroom, during consulting sessions, at advisory board meetings, etc. Perhaps model users of the future will demand model validation information, and they will know how to use it.

TABLE 1

List of basic components to be considered when assessing the validity and uncertainty of a mathematical model

INPUT
- Variables
  - Selection
  - Surrogate or direct
  - Quality of measurements
- Parameters
  - Selection and model identifiability
  - Calibration (parameter estimation) methodology
    - Design for acquiring calibration data
    - Estimation technique
    - Assumptions and Diagnostics
- Side conditions and constraints
- Subjective input and expert judgment

STRUCTURE
- Face validity and plausibility
- Numerical validity
  - Approximation properties (due to simplifying, lumping, linearizing, etc.)
  - Mathematical properties (internal validity)
- Computer programs &/or algorithms
- Sensitivity of model and submodels
- Limitations, boundary conditions, etc.

OUTPUT
- Variables
  - Selection
  - Surrogate variables or direct variables
  - Interpolations or extrapolations
  - Quality of measurements
- Data sets
  - Design considerations
  - Assessment of dependencies (spatial correlation, serial correlation, etc.)
  - Evaluation strategy
    - New, independent evaluation data
    - Hold-out data independent of data used for model calibration, etc.
    - Other: cross-validation, jackknifing, bootstrapping, etc.
- Statistical evaluation
  - Statistics (for measuring forecasting accuracy; goodness-of-fit; etc.)
    - For assessment relative to other models (e.g., ranking and selection)
    - For assessment relative to established requirements for acceptability
    - For describing validity (appropriate for a specific use of the model)
  - Methodology
    - Assumptions and diagnostics
- Non-statistical evaluation

BIBLIOGRAPHY

1. Ahmad, M.I., Sinclair, C.D. & Spurr, B.D. (1988). Assessment of flood frequency models using empirical distribution function statistics. *Water Resources Research* **24**, 1323-1328.
Modifies the Anderson-Darling goodness-of-fit test to be sensitive to departures in the upper (or lower) tail of the distribution. Critical values are found by Monte Carlo methods.
Keywords: goodness-of-fit, statistics, extreme events, water research

2. Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* **46**, 293-302.
Proposes a measure of agreement between two classification procedures. Derives estimation and inference methods.
Keywords: statistics, goodness-of-fit, prediction measures, output, classification, uncertainty, biomedical

3. Alatiqi, I.M., Meziou, A.M. & Gasmelseed, G.A. (1989). Modelling, simulation and sensitivity analysis of steam-methane reformers. *International Journal of Hydrogen Energy* **14**, 241-256.
Evaluates two models against actual data and for face validity. Models are for supervisory and automatic control design.
Keywords: mechanistic, control, engineering, structure, prediction, prediction measures, face validity

4. Alcamo, J. & Bartnicki, J. (1987). A framework for error analysis of a long-range transport model with emphasis on parameter uncertainty. *Atmospheric Environment* **21**, 2121-2131.
Proposes steps to follow in assessing a model. Uses Monte Carlo methods to investigate model sensitivity. Itemizes some difficulties in estimating the vector of true errors between model predicted output and true output.
Keywords: air pollution, input, output, sensitivity, mechanistic, uncertainty

5. Amaral, D. (1988). Including uncertainty in assessments of sulfur oxide health risks. *Journal of the Air Pollution Control Association* **38**, 399-405.
Describes experts' judgment concerning health risks from a hypothetical new coal-fired power plant. The experts' predicted number of deaths per year ranged from zero to thousands.
Keywords: air pollution, uncertainty, prediction, expert opinion

6. Anderson, S.L. (1990). Random number generators on vector supercomputers and other advanced architectures. *SIAM Review* **32**, 221-251.
Discusses methods of testing goodness-of-fit to the uniform distribution, including plots as well as statistical tests.
Keywords: statistics, testing, goodness-of-fit, graphics

7. Andriole, S.J. (Editor) (1986). *Software Validation, Verification, Testing, and Documentation*, Princeton, NJ:Pertocelli Books.
This source book was edited from a number of government reports produced by the National Bureau of Standards. Each of the five sections contains a glossary. Recommendations are included; e.g., a recommended table of contents for a users manual and a checklist of information required for model documentation.
Keywords: verification, testing, documentation, software

8. Anthes, R.A. (1983). Regional models of the atmosphere in middle latitudes. *Monthly Weather Review* **111**, 1306-1335.
Reviews numerical methods for weather prediction over limited- area, regional domains. Discusses numerical and physical factors important to achieving prediction accuracy. Discusses many measures of forecast skill and a wide variety of plots.
Keywords: structure, forecasting, prediction measures, numerical analysis, input, mechanistic, meteorology, graphics

9. Apostolakis, G. (1990). The concept of probability in safety assessments of technological systems. *Science* **250**, 1359-1364.
Argues that mathematical models, expert opinions, and Bayesian analyses are necessary in order to assess the risk of complex engineering systems such as containment of radioactive waste repositories. Good reference list.
Keywords: energy, engineering, environmental, expert opinion, Bayesian, uncertainty

10. Arlat, J., Aguera, M., Amat, L., Crouzet, Y., Fabre, J., Laprie, J., Martins, E. & Powell, D. (1990). Fault injection for dependability validation: A methodology and some applications. *IEEE Transactions on Software Engineering* **16**, 166-182.
Discusses fault injection, the deliberate introduction of input or of hardware faults into the program, to determine the fault tolerant capabilities of the program.
Keywords: engineering, design, software

11. Armstrong, J.S. (1988). Research needs in forecasting. *International Journal of Forecasting* **4**, 449-465.
Describes the results of various surveys concerning forecast techniques. Discusses audits of forecasting models and the legal responsibilities of forecasters.
Keywords: review, forecasting, general

12. Arratia, R., Gordon, L. & Waterman, M.S. (1990). The Erdos-Renyi law in distribution, for coin tossing and sequence matching. *Annals of Statistics* **18**, 539-570.
Derives an approximate tail probability for the number of matching regions in strands of DNA from two organisms. Compares the theoretical calculations to simulated realizations using an empirical histogram hanging from the predicted histogram after arc-sine transformation.
Keywords: theoretical biology, statistics, testing, goodness-of-fit, graphics

13. Atkinson, A.C. (1986). Regression diagnostics. In *Encyclopedia of statistical sciences - Volume 7*, Eds. Kotz, S., Johnson, N.L. & Read, C.B., pp. 689-694. New York, New York: John Wiley & Sons.
Brief review of techniques for detecting disagreement between a multiple linear regression model and the data to which it is fitted.
Keywords: statistics, regression, diagnostics, graphics, empirical, calibration

14. Atwood, C.L. (1986). Correct Bayesian and frequentist intervals are similar. *Nuclear Engineering and Design* **93**, 145-151.
Reviews articles concerning expert judgment and the uncertainties experts attach to those judgments. Concludes that experts may be quite wrong in their beliefs, disagree greatly with each other, and cannot measure their own inaccuracies well. Is very skeptical of informative priors in Bayesian analyses.
Keywords: statistics, Bayesian, expert opinion, uncertainty, input, structure, energy, review

15. Azzalini, A., Bowman, A.W. & Hardle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1-11.
Proposes the use of nonparametric regression and resampling simulation techniques to ascertain the goodness-of-fit of a parametric model. Illustrated for binary response data, Poisson regression, and time series data.
Keywords: structure, output, graphics, simulation, logistic regression, serial correlation, time series, regression, statistics, uncertainty, cross-validation, diagnostics, goodness-of-fit

16. Badescu, V. (1988). Comment on the statistical indicators used to evaluate the accuracy of solar radiation computing models. *Solar Energy* **40**, 479-480.
Defines some statistical measures commonly used for model evaluation.
Keywords: output, prediction, prediction measures, meteorology

17. Balci, O. (1989). How to assess the acceptability and credibility of simulation results. *Proceedings of the 1989 Winter Simulation Conference* **1**, 62-71.
Presents a comprehensive life cycle of a computer simulation study and lists 10 processes, 10 phases, and 13 credibility assessment stages of the life cycle. The lists could be used to guide the investigator through a complete model validation routine, beginning at the conceptual stage and ending with the documentation stage.
Keywords: simulation, software, review, documentation

18. Balci, O. & Sargent, R.G. (1981). A methodology for cost-risk analysis in the statistical validation of simulation models. *Communications of the ACM* **24**, 190-197.
Discusses testing the null hypothesis that the simulation model is valid for the acceptable range of accuracy. Considers design selection for collecting data to achieve desired power for this test.
Keywords: statistics, output, testing, engineering, design

19. Balgovind, R., Dalcher, A., Ghil, M. & Kalnay, E. (1983). A stochastic-dynamic model for the spatial structure of forecast error statistics. *Monthly Weather Review* **111**, 701-722.
Suggests models for the spatially two-dimensional, latitude- dependent structure of forecast error correlations where the purpose is to interpolate within observation points not uniformly distributed in space.
Keywords: time series, serial correlation, structure, dynamic, spatial model, mechanistic

20. Barinage, M. (1990). Neuroscience models the brain. *Science* **247**, 524-526.
Describes status of mathematical models for brain function and how such models are used to guide experimentation
Keywords: theoretical biology, discovery, explanation

21. Baskerville, J.C. & Toogood, J.H. (1982). Guided regression modeling for prediction and exploration of structure with many explanatory variables. *Technometrics* **24**, 9-17.
An overview of multiple linear regression modeling, including diagnostics and variable selection techniques. Describes a principle component analysis approach to variable selection.
Keywords: statistics, structure, empirical, output, prediction, sensitivity, graphics, diagnostics, selection, regression, explanation

22. Bates, D.M. & Hunter, W.G. (1985). Nonlinear models. In *Encyclopedia of statistical sciences - Volume 6*, Eds. Kotz, S., Johnson, N.L. & Read, C.B., pp. 293-301. New York, New York: John Wiley & Sons.
Review of nonlinear models with an extensive bibliography. Information on nonlinear least squares regression, multiresponse estimation, design of experiments, and diagnostics.
Keywords: structure, statistics, nonlinear models, calibration, bibliography

23. Baveye, P. & Valicchi, A. (1989). An evaluation of mathematical models of the transport of biologically reacting solutes in saturated soils and aquifers. *Water Resources Research* **25**, 1413-1421.
Three models produced similar predictions. Assessment focused on interpretation of parameters and physical validity of substructures.
Keywords: water research, face validity, substructure, calibration, prediction, mechanistic, theoretical biology

24. Beck, M.B. (1986). The selection of structure in models of environmental systems. *The Statistician* **35**, 151-161.
Starting with a simple mechanistic model for a time-series of input-output data, the author re-estimates the parameters of the model as each successive time point is added to the data set. Then a plot of parameter estimates versus time, with an error bound around the estimate, shows whether the parameter is staying relatively constant. Provides a test of the stationary assumption.
Keywords: structure, statistics, mechanistic, time series, empirical, selection, testing, stationary

25. Beck, M.B. & van Straten, G.(Eds.) (1983). *Uncertainty and forecasting of water quality*, New York:Springer-Verlag.
In the opening chapter, Beck discusses methods for investigating uncertainties in hypothesized structures and in predictions. For the former, he advocates recursive estimation algorithms and experimentation. McLaughlin describes propagation of error based on first order Taylor expansions. Gardner and O'Neill review issues in model sensitivity analyses and describes some advantages of simulation techniques. Young gives an overview of model assessment, with discussion of the terminology. Sharefkin suggests generating pseudodata by one of the models and seeing how each model performs on the pseudodata. Possibly, one model is satisfactory, regardless of which of the models is correct. Of course, this does not determine which model best fits reality.
Keywords: water research, economics, Bayesian, prediction, sensitivity, face validity, uncertainty, forecasting, sensitivity measures, calibration, design, expert opinion, mechanistic, philosophy

26. Beizer, B. (1984). *Software System Testing and Quality Assurance*, New York:Van Nostrand Reinhold Company.
Brief overview of software quality assurance with definitions of terms.
Keywords: quality control, software

27. Belsley, D.A. (1986). Model selection in regression analysis, regression diagnostics and prior knowledge. *International Journal of Forecasting* **2**, 41-52.
Reviews diagnostics and selection schemes and describes some dangers of diagnostics which lead to casting out observations or transforming the model, especially when the goal is extrapolation. Followed by interesting and useful discussion.
Keywords: statistics, forecasting, prediction, structure, output, diagnostics, empirical, mechanistic, selection, regression, Bayesian, review

28. Belsley, D.A. (1988). Modelling and forecasting reliability. *International Journal of Forecasting* **4**, 427-447.
Argues that a mechanistic model is required if the model is intended to be anything more than descriptive. Discusses interpretations of the deterministic and stochastic components of a model and the role of model testing.
Keywords: economics, forecasting, testing, mechanistic, structure, prediction, output

29. Belsley, D.A. & Kuh, E. (1986). *Model Reliability*, Cambridge, MA:Massachusetts Institute of Technology Press.
Collection of papers discussing the inadequacy of existing models and the fragility of their predictions.
Keywords: economics, forecasting, general

30. Belsley, D.A., Kuh, E. & Welsch, R.E. (1980). *Regression Diagnostics*, New York:John Wiley & Sons, Inc..
This comprehensive book discusses diagnostics and remedies for influential observations, outliers, and collinearity in multiple linear regression. Bayesian methods are mentioned.
Keywords: statistics, regression, diagnostics, graphics, empirical, Bayesian, residual analysis

31. Bencala, K.E. & Seinfeld, J.H. (1979). An air quality model performance assessment package. *Atmospheric Environment* **13**, 1181-1185.
Describes a long list of methods for measuring the ability of the model to fit observed data and announces the availability of a computer package to do the calculations. Techniques for dealing with serial or spatial correlations are not mentioned.
Keywords: air pollution, output, prediction measures, prediction, computer programs

32. Bendell, T. (1986). Minimizing misconceived models. *The Statistician* **35**, 303-309.
Provides an example where two different underlying mechanisms generate the same negative binomial model.
Keywords: structure

33. Beran, J. (1989). A test of location for data with slowly decaying serial correlations. *Biometrika* **76**, 261-269.
Proposes a modified t-test for a location parameter which takes into account slowly decaying serial correlations.
Keywords: statistics, serial correlation, spatial correlation, time series, testing, output, hydrology

34. Berk, K.N. (1984). Validating regression procedures with new data. *Technometrics* **26**, 331-338.
Concerns the use of hold-out data for evaluating empirical models. Shows how the MSE for the evaluation set may well be higher than the MSE for the estimation set, even if the model were exactly correct for both sets.
Keywords: statistics, prediction, regression, selection, hold-out data, output, empirical

35. Besag, J. & Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* **76**, 633-642.
Generalized Monte Carlo testing methods can be used to test the association between two sets of observations taken at a variety of spatial locations. Highly mathematical.
Keywords: statistics, testing, output, prediction, spatial correlation, simulation, permutation test

36. Beven, K. (1989). Changing ideas in hydrology - the case of physically-based models. *Journal of Hydrology* **105**, 157-172.
Criticizes the application of mechanistic models for practical prediction in hydrology. Discusses effects of lumping, overparameterization, identifiability, and uncertainties.
Keywords: mechanistic, hydrology, uncertainty

37. Bloch, D.A. (1990). Evaluating predictions of events with binary outcomes: An appraisal of the Brier score and some of its close relatives. Technical Report No. 135, Division of Biostatistics, Stanford University, Stanford, CA. (UnPub)
Deals with two problems: evaluating the accuracy of a single set of predictions of binary outcomes and comparing the predictive accuracy of two sets of predictions for the same binary outcomes.
Keywords: statistics, prediction, prediction measures, forecasting, testing, permutation test, output

38. Booker, J.M. (1988). Combining qualitative and quantitative data: the use of subjective judgment modeling; Report LA-UR-88-1675, Los Alamos National Laboratory, Los Alamos, NM. (UnPub)
Issues to be considered when expert judgment is to entered into a model, including granularity, conditionality, and quantification. Reviews methods for eliciting expert opinions.
Keywords: statistics, expert opinion, structure

39. Borgman, L.E. (1988). New advances in methodology for statistical tests useful in geostatistical studies. *Mathematical Geology* **20**, 383-403.
Generalizes the sign, Wilcoxon, and some parametric tests for application to correlated random fields (correlation function known).
Keywords: spatial correlation, testing, output, prediction measures, statistics

40. Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A* **143**, 383-430.
General discussion focusing on the relative contributions of frequentist and Bayesian approaches. Shows that the model can be factored into a Bayesian part and a predictive part. With discussion and rejoinder.
Keywords: statistics, Bayesian, diagnostics, serial correlation, outliers, prediction, prediction measures, predictive distributions, testing, uncertainty

41. Box, G.E.P. (1984). The importance of practice in the development of statistics. *Technometrics* **26**, 1-8.
Reviews important developments in statistics and divides problems into those for which relative frequency approaches are better and those for which Bayesian approaches are better. Discusses the scientific method, model building, model modification, analysis of residuals, tests of fit, and diagnostic checks, both graphical and numerical.
Keywords: statistics, philosophy, Bayesian, graphics, diagnostics, historical, explanation

42. Box, G.E.P. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics* **30**, 1-40.
Critically evaluates the Taguchi performance criteria based on signal-to-noise ratios. Proposes alternative procedures based on conventional statistical design and analysis principles. With discussion.
Keywords: statistics, quality control, design, graphics, testing, output, prediction measures

43. Box, G.E.P. & Draper, N.R. (1987). *Empirical Model-building and Response Surfaces*, New York:John Wiley & Sons, Inc..
This popular text book contains a general discussion of modelling, especially as used in science. The book focuses on empirical response surface modeling over a specific region of the predictor space and the design of the experiments that are to provide data.
Keywords: statistics, structure, empirical, prediction, control

44. Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*, New York:John Wiley & Sons, Inc..
Describes the steps required in model-building. Advocates Bayesian approaches to model selection. Discusses designing experiments for model selection, model testing, methods of parameter estimation, and visual diagnostic techniques.
Keywords: statistics, empirical, mechanistic, calibration, sensitivity, design, diagnostics, graphics

45. Buckland, S.T. (1985). Calculation of Monte Carlo confidence intervals. *Applied Statistics* 34, 296-301.
A FORTRAN subroutine for calculating bootstrap confidence intervals. Potentially useful for ascertaining the uncertainty of prediction measures.
Keywords: statistics, bootstrap, computer programs

46. Bush, R.R. & Mosteller, F. (1988). A comparison of eight models. In *Readings in mathematical social science*, Eds. Lazarsfeld, P.F. & Henry, N.W., pp. 335-349. Cambridge, Massachusetts: The M.I.T. Press.
Describes a simulation experiment in which output data are simulated from 8 different models and each data set is compared to real experimental data according to a variety of criteria.
Keywords: behavioral science, output, prediction measures, prediction, substructure, testing

47. Buxton, B.E.(Ed.) (1987). *Geostatistical, Sensitivity, and Uncertainty Methods for Ground-water Flow and Radionuclide Transport Modeling*, Columbus, OH:Battelle Press.
Eisenberg, Rickertsen, and Voss discuss major sources of uncertainty and look at some specific examples; e.g., characterizing ground-water travel time. Brandstetter and Buxton discuss performance assessment, which is very similar to model assessment, but it has the added element of uncertainty due to scenario (the range and distribution of appropriate input variables and parameters and exogenous factors that can alter the model). They provide an outline of geostatistical methods, sensitivity analysis, and uncertainty analysis. Worley, Oblow, Pin, Maerker, Horwedel, Wright, and Lucius describe two automatic derivative calculation and error propagation computer programs for use in sensitivity analyses. They favor a deterministic uncertainty analysis technique based on approximating the response surface with a model that incorporates knowledge of the partial derivatives of the true response with respect to the parameters of interest. Nicholson, Wierenga, Gee, Jacobson, Polmann, McLaughlin, and Gelhar propose a validation protocol that begins with a preliminary field experiment to estimate parameter values and to determine whether the parameter is spatially varying. Then a large scale field experiment is conducted and the model is run and evaluated. Merkhofer and Runchal show that experts trained in probability assessment are typically surprised between 5 and 50 percent of the time. They discuss various reasons for experts' excess confidence, and proposes approaches for helping experts produce better assessments.
Keywords: water research, spatial model, uncertainty, Bayesian, expert opinion, stochastic, deterministic, mechanistic, testing, output, review, computer programs, performance analysis, forecasting

48. Bye, B.V. & Dykacz, J.M. (1987). The effects of misspecification in logistic regression models. *American Statistical Association Proceedings of the Social Statistics Section* 98-103.
Presents the results of a computer simulation study of misspecification in logistic multiple regression when the predictors are observed, not chosen by design.
Keywords: statistics, logistic regression, misspecification, empirical

49. Calver, A. (1988). Calibration, sensitivity and validation of a physically-based rainfall-runoff model. *Journal of Hydrology* **103**, 103-115.
This paper describes the calibration, sensitivity analysis, checks of face validity, and assessment against data of a rainfall-runoff model. Cautioned that the model is physically--based rather than entirely physically descriptive.
Keywords: hydrology, calibration, sensitivity, output, prediction, prediction measures, face validity

50. Campbell, K. (1988). Bootstrapped models for intrinsic random functions. *Mathematical Geology* **20**, 699-715.
Proposes a bootstrap method for determining the variability or confidence interval when kriging.
Keywords: spatial model, statistics, bootstrap, uncertainty, output, testing, earth science

51. Carbone, R., Andersen, A., Corriveau, Y. & Corson, P.P. (1983). Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Science* **29**, 559-565.
A forecasting competition was conducted using 25 time series of data. Concludes that technical expertise, judgmental adjustment, and individualized analyses were of little value in improving forecast accuracy and that simpler methods (exponential smoothing) provided significantly more accurate forecasts than the Box-Jenkins method when applied by persons with limited training.
Keywords: forecasting, time series, prediction, empirical

52. Carbone, R. & Armstrong, J.S. (1982). Evaluation of extrapolative forecasting methods: results of a survey of academicians and practitioners. *Journal of Forecasting* **1**, 215-217.
In an informal survey, experts in the field were asked to list the important criteria to consider when evaluating forecasting methods. Among those who listed any statistical measures of accuracy, the mean squared error was most often mentioned.
Keywords: output, prediction, prediction measures, forecasting,

53. Carroll, R.J. & Ruppert, D. (1988). *Transformations and Weighting in Regression*, New York:Chapman & Hall.
Discusses estimation, graphical methods, robustness, selection, diagnostics, and uncertainty calculations when the error term in the regression model has an unknown variance function.
Keywords: regression, statistics, diagnostics, selection, graphics, calibration, uncertainty, empirical

54. Caswell, H. (1977). The validation problem. In *Systems Analysis and Simulation in Ecology - Volume IV*, Ed. Patten, B.C., pp. 313-325. New York: Academic Press.
A general overview of model validation. Differentiates between models for science and models for prediction. Uses Popper's corroboration for scientific models.
Keywords: ecology, philosophy, structure, output, prediction, mechanistic, empirical

55. Chatfield, C. (1988). What is the 'best' method of forecasting?. *Journal of Applied Statistics* **15**, 19-38.
Reviews forecasting methods, which are classified as univariate, multivariate, or judgmental, and also as automatic or non-automatic. Recommends forecast monitoring.
Keywords: economics, statistics, forecasting, prediction measures, time series, structure, output, surveillance

56. Chatterjee, S. & Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression with comments and reply. *Statistical Science* **1**, 379-416.
Reviews regression diagnostics for discovering single points that are influential, of high leverage, or outliers. The authors and discussants do not agree on the relative merits of the various measures, but might agree on a basic checklist provided in the discussion by Hoaglin and Kempthorne. The consensus favors graphical diagnosis. Some real data sets are listed and analyzed. Some very informative plots of contrived data are presented to illustrate various measures.
Keywords: structure, output, statistics, regression, empirical, diagnostics, outliers, selection, graphics, review, residual analysis

57. Chatterjee, S. & Hadi, A.S. (1988). *Sensitivity analysis in linear regression*, New York:John Wiley & Sons.
An excellent, modern review of diagnostic techniques in multiple linear regression. It contains theory, methods, and examples of techniques for discovering sets of outliers, influential observations, sets of high leverage, testing goodness-of-fit, etc.
Keywords: statistics, regression, diagnostics, sensitivity, structure, calibration, graphics, review

58. Christensen, R. (1989). Lack-of-fit tests based on near or exact replications. *Annals of Statistics* **17**, 673-683.
Derives an exact version of the usual F test for model fit based on a pure error denominator.
Keywords: statistics, regression, diagnostics, empirical, structure, testing

59. Clemen, R.T. & Winkler, R.L. (1987). Calibrating and combining precipitation probability forecasts. In *Probability and Bayesian Statistics*, Ed. Viertl, R., pp. 97-110. New York: Plenum Press.
After trying some complicated methods for combining probability of precipitation forecasts, the authors concluded that a simple average of forecasts does better than more complex methods.
Keywords: forecasting, Bayesian, prediction, meteorology, prediction measures

60. Cliff, A.D. & Ord, J.K. (1981). *Spatial Processes: Models and Applications*, London:Pion Limited.
Discusses the application of standard statistical methods, such as the t-test and linear model analyses, to cases where spatial correlations are present. Contains a nice review of the literature and a useful reference list.
Keywords: earth science, structure, empirical, time series, serial correlation, bibliography, spatial correlation, spatial model

61. Clifford, P., Richardson, S. & Hemon, D. (1989). Assessing the significance of the correlation between two spatial processes. *Biometrics* **45**, 123-134.
Associated with each observed and predicted multivariate output is a time and space location. Discusses approaches to testing the null hypothesis of no correlation between X and Y.
Keywords: spatial correlation, testing, output, matching, spatial model

62. Cobelli, C., Carson, E.R., Finkelstein, L. & Leaning, M.S. (1984). Validation of simple and complex models in physiology and medicine. *American Journal of Physiology* **246**, R259-R266.

Proposes a vocabulary for modeling and model validation. Describes the steps required in assessing a model. Good overview.

Keywords: biomedical, input, output, structure, mechanistic, sensitivity, theoretical biology, general, calibration

63. Cobelli, C. & DiStefano, J.J. III (1980). Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology* **239**, R7-R24.

Excellent survey of identifiability of parameters when a model is being calibrated. The focus is on deterministic models, especially compartment models. Two difficulties are that, even if a system is identifiable, there is no guarantee that a unique set of parameter values can be determined, and, even if a unique set of parameter values exists, there may not be a known method for calculating them.

Keywords: mathematics, mechanistic, numerical analysis, structure, calibration, biomedical, deterministic, design, dynamic, testing, theoretical biology, tracer

64. COHMAP, M.embers. (1988). Climatic changes of the last 18,000 years: observations and model simulations. *Science* **241**, 1043-1052.

The authors have assembled data that provide geologic records of spatial and temporal changes in climate from radiocarbon- dated stratigraphic sequences from several continents and oceans. These data are compared with model simulations of past climates, aggregated over intervals as wide as 3000 years. The data-model comparisons are visual.

Keywords: prediction, mechanistic, graphics, simulation, dynamic, earth science

65. Collings, B.J. & Hamilton, M.A. (1988). Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics* **44**, 847-860.

Shows how to use the bootstrap technique to determine the power of two-sample distribution-free tests.

Keywords: statistics, output, prediction, testing, bootstrap, calibration

66. Cook, D.G. & Pocock, S.J. (1983). Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics* **39**, 361-371.

Suggests a two-step procedure to account for spatial correlations among the cases when fitting a multiple linear regression model. Applied to cardiovascular mortality in British towns.

Keywords: statistics, spatial correlation, regression, testing, uncertainty, structure, biomedical, empirical

67. Cook, R.D., Tsai, C.L. & Wei, B.C. (1986). Bias in nonlinear regression. *Biometrika* **73**, 615-623.

Describes diagnostics useful in nonlinear least squares regression analysis.

Keywords: nonlinear regression, structure, mechanistic, statistics, empirical, diagnostics, graphics, testing

68. Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*, New York:Chapman and Hall.

A statistics textbook presentation of methods of regression diagnostics.

Keywords: statistics, regression, diagnostics, empirical, graphics, residual analysis

69. Cooley, R.L., Konikow, L.F. & Naff, R.L. (1986). Nonlinear-regression groundwater flow modeling of a deep regional aquifer system. *Water Resources Research* **22**, 1759-1778.

Alternative mechanistic models are fit by nonlinear regression and compared pairwise using two statistics, the ratio of residual variances and an approximate F statistic. A sensitivity analysis was conducted, and estimated parameters were compared to published values.

Keywords: structure, output, hydraulics, selection, sensitivity, calibration, dynamic, mechanistic, numerical model, nonlinear regression

70. Copas, J.B. (1987). Cross-validation shrinkage of regression predictors. *Journal of the Royal Statistical Society, Series B* **49**, 175-183.

Shows how to use cross-validation to determine the amount of shrinkage. Examples and a simulation study are presented.

Keywords: cross-validation, regression, prediction, output, calibration, empirical, prediction measures

71. Copas, J.B. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society, Series B* **50**, 225-265.

Interesting review and extension of logistic regression methodology with a focus on detecting or accommodating outliers. With discussion.

Keywords: logistic regression, statistics, diagnostics, empirical

72. Costanza, R. & Sklar, F.H. (1985). Articulation, accuracy and effectiveness of mathematical models: A review of freshwater wetland applications. *Ecological Modelling* **27**, 45-68.

The authors build indices of mathematical complexity, descriptive accuracy, degree of articulation, and effectiveness. Then they evaluate dozens of models by these criteria. This review suggests the beginnings of a theory of modelling.

Keywords: ecology, prediction, prediction measures, mechanistic, review,

73. Cox, D.R. (1977). The role of significance tests. *Scandinavian Journal of Statistics* **4**, 49-70.

Sets out those circumstances under which significance tests are likely to be valuable in the intermediate stages of an analysis or in the final statement of conclusions. With discussion.

Keywords: statistics, testing, diagnostics, goodness-of-fit, selection, specification

74. Cressie, N. (1989). Geostatistics. *American Statistician* **43**, 197-202.

Surveys the geostatistical techniques of variograms and kriging, illustrated on spatial data in two dimensions.

Keywords: spatial model, statistics, prediction, stationary, empirical, earth science, review

75. Cressie, N. & Read, R.C. (1989). Pearson's χ^2 and the log-likelihood ratio statistic G^2: a comparative review. *International Statistical Review* **57**, 19-43.

Because power-divergence statistics are a link between the Pearson's chi-squared and the loglikelihood ratio statistics, they may well lead to new goodness-of-fit statistics.

Keywords: statistics, goodness-of-fit, testing, serial correlation, historical, review

76. Creutin, J.D. & Obled, C. (1982). Objective analyses and mapping techniques for rainfall fields: An objective comparison. *Water Resources Research* **18**, 413-431.

Reviews methods for estimating point or average rainfall values at ungauged sites. Among the methods are nearest neighbor, spline-surface fitting, optimal interpolation, kriging, and surface fitting with empirical orthogonal functions. Recommends optimal interpolation.

Keywords: water research, review, spatial model, spatial correlation, empirical, cross-validation, hold-out data, graphics, prediction measures

77. Crowder, H., Dembo, R.S. & Mulvey, J.M. (1979). On reporting computational experiments with mathematical software. *ACM Transactions on Mathematical Software* **5**, 193-203.
Thorough discussion, with recommendations, of assessing mathematical algorithms and evaluating the efficiency of competing computer codes.
Keywords: mathematics, structure, numerical analysis, substructure, software

78. Cunningham, A.B. & Sinclair, P.J. (1979). Application and analysis of a coupled surface and groundwater model. *Journal of Hydrology* **43**, 129-148.
Emphasis is placed on evaluation of model performance rather than on model development. Lists the various sources of uncertainty involved with modeling a real hydraulic system. Discusses analyses that focus on model sensitivity, identifiability, and a comparison of the predicted and observed time series (not specifically accounting for potential serial correlations).
Keywords: hydrology, prediction, output, sensitivity, serial correlation, mathematics, numerical analysis

79. D'Agostino, R.B. & Stephens, M.A. (1986). *Goodness-of-fit Techniques*, New York:- Marcel Dekker, Inc..
A readable, usable compilation of methods for assessing the fit of a particular distributional shape to a single sample of data, including these approaches: plotting, chi-squared, empirical distribution function, regression and correlation, transformations, and moments. Computational details and numerical examples are provided. Whenever possible, recommendations and rules-of-thumb are listed.
Keywords: goodness-of-fit, statistics, structure

80. Dagan, G. & Bresler, E. (1988). Variability of yield of an irrigated crop and its causes 3: Numerical simulation and field results. *Water Resources Research* **24**, 395-401.
A model for predicting crop yield was compared to actual yield and a sensitivity analysis was conducted.
Keywords: agriculture, water research, calibration, sensitivity, output, prediction, prediction measures

81. Davidian, M. & Carroll, R.J. (1987). Variance function estimation. *Journal of the American Statistical Association* **82**, 1079-1091.
Reviews methods for estimating the variance function with emphasis on robust techniques.
Keywords: output, prediction, statistics, regression, weighted regression, biomedical, control, uncertainty, empirical

82. Davis, P.A., Reimer, A., Sakiyama, S.K. & Slawson, P.R. (1986). Short-range atmospheric dispersion over a heterogeneous surface-I Lateral dispersion. *Atmospheric Environment* **20**, 41-50.
Use data from tracer releases to study cross-wind concentration distributions, especially the standard deviation and the extent to which the shape was approximated by a Gaussian distribution. Mechanistic and empirical models were used to predict the standard deviation.
Keywords: substructure, mechanistic, empirical, meteorology, prediction measures, input, tracer

83. Dawes, R.M., Faust, D. & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science* **243**, 1668-1674.
Reviews the results of a variety of studies comparing empirical methods of clinical diagnosis, using regression or discriminant analysis, to diagnosis by experts. In-depth discussion of the

advantages and disadvantages of empirical methods. Subsequent discussion in Science(1990) 247:146-147.
Keywords: empirical, behavioral science, expert opinion, cross-validation, hold-out data

84. de Silans, A.P., Bruckler, L., Thony, J.L. & Vauclin, M. (1989). Numerical modeling of coupled heat and water flows during drying in a stratified bare soil - comparison with field observations. *Journal of Hydrology* **105**, 109-138.
Discusses calibration, sensitivity analysis, and comparison of predictions to post-calibrations data.
Keywords: hydrology, calibration, sensitivity, output, prediction

85. Deaton, M.L. (1988). Simulation models, validation of. In *Encyclopedia of statistical sciences - Volume 8*, Eds. Kotz, S., Johnson, N.L. & Read, C.B., pp. 481-484. New York: John Wiley & Sons.
Provides a brief overview of practical difficulties in model validation and of statistical methods for evaluating validity.
Keywords: statistics, general, simulation

86. DeGroot, M.H. (1988). A Bayesian view of assessing uncertainty and comparing expert opinion. *Journal of statistical planning and inference* **20**, 295-306.
Among well-calibrated experts, the best experts are those whose predictions are most variable.
Keywords: expert opinion, Bayesian, statistics, forecasting, prediction, prediction measures

87. Demerjian, K.L. (1985). Quantifying uncertainty in long-range-transport models: a summary of the AMS workshop on sources and evaluation of uncertainty in long-range-transport models. *Bulletin of the American Meteorological Society* **66**, 1533-1540.
Reviews the AMS/EPA Workshop, Woods Hole, MA, September, 1984, which addressed the quantification of uncertainty in long-range transport (LRT) model predictions. The working group was especially concerned with the analysis and interpretation of available input data, with methods of parameterization, with some of the substructure components of the models, and with missing substructure components.
Keywords: data quality, meteorology, sensitivity, air pollution, prediction, input, calibration, substructure

88. den Butter, F.A.G. & Verbon, H.A.A. (1982). The specification problem in regression analysis. *International Statistical Review* **50**, 267-283.
Describe a modelling strategy comprising the identification stage, the estimation stage, and the stage of diagnostic checking. Reviews the specification arguments for a number of macro-economic models, many of which were originally postulated with neither economic nor statistical motivation.
Keywords: economics, selection, regression, logistic regression, empirical, structure, serial correlation

89. Dennis, R.L. (1985). Issues, design and interpretation of performance evaluations: ensuring the emperor has clothes. *Preprints Volume, 15th Inter. Tech. Meeting on Air Pollution Modeling and Its Applications* V.2, 1-13.
Considers the similarity of the distribution of predictions to the distribution of observations. Discusses the prediction of daily maximum concentrations.
Keywords: meteorology, graphics, output, prediction, air pollution, testing, goodness-of-fit, performance analysis

90. Dennis, R.L. & Downton, M.W. (1984). Evaluation of urban photochemical models for regulatory use. *Atmospheric Environment* **18**, 2055-2069.
Reviews the use and interpretation of statistical measures for urban air quality model performance in predicting daily maximum concentrations. Autocorrelations in the differences (predicted-observed) were accounted for using the method of Downton and Dennis (1985). Tries alternative schemes for pairing predicted with observed.
Keywords: air pollution, output, prediction measures, prediction, mechanistic, extreme events

91. Deshpande, J.V. & Mehta, G.P. (1983). Nonparametic procedures to select populations better than a known standard. *Sankhya: the Indian Journal of Statistics, Series B* **45**, 330-344.
Discusses a nonparametric subset selection scheme.
Keywords: statistics, ranking and selection, testing

92. Diaconis, P. & Mosteller, F. (1989). Methods for Studying Coincidences. *Journal of the American Statistical Association* 84, 853-861.
This important article illustrates basic statistical techniques for separating coincidences from potential cause-effect relationships.
Keywords: philosophy, input, statistics, structure, causal effects, explanation

93. DiCiccio, T.J. & Romano, J.P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society, Series B* **50**, 338-354.
Discusses several distinct bootstrap methods for constructing confidence regions with emphasis on the mathematical correctness of the procedures. Paired with Hinkley (1988) with discussion of both.
Keywords: bootstrap, statistics, output, uncertainty

94. DiCiccio, T.J. & Romano, J.P. (1990). Nonparametic confidence limits by resampling methods and least favorable families. *International Statistical Review* **58**, 59-76.
This mathematical paper synthesizes several bootstrap approaches by use of the notion of a least favorable distribution. Large sample properties are given.
Keywords: statistics, bootstrap, uncertainty

95. Dielman, T.E. & Pfaffenberger, R.C. (1989). Efficiency of ordinary least squares for linear models with autocorrelation. *Journal of the American Statistical Association* **84**, 248-248.
Shows that Kramer's (1980) results understate the advantages of correcting for autocorrelation when conducting a linear regression analysis.
Keywords: statistics, empirical, serial correlation, calibration

96. Dingman, J.S. & Bedford, K.W. (1986). Skill tests and parametric statistics for model evaluation. *Journal of Hydraulic Engineering-ASCE* **112**, 124-141.
Presents and evaluates a variety of traditional statistical methods and nonparametric skill tests for their abilities to determine the accuracy of surface water models in reproducing the measured time and space histories of calculated variables.
Keywords: output, prediction, prediction measures, extreme events, mechanistic, water research, engineering, hydraulics

97. Docter, R., deJong, M., van der Hoek, H.J., Krenning, E.P. & Hennemann, G. (1990). Development and use of a mathematical two-pool model of distribution and metabolism of 3,3',5-triiodothyronine in a recirculating rat liver perfusion system: Albumin does not play a role in cellular transport. *Endocrinology* **126**, 451-459.
A two-compartment model, fit by an ad hoc application of least squares techniques, was used to explain the kinetics of the process.
Keywords: theoretical biology, mechanistic, output, prediction measures, calibration, explanation

98. Doktor, R.H. & Chandler, S.M. (1988). Limits of predictiability in forecasting in the behavioral sciences. *International Journal of Forecasting* **40**, 5-14.
Describes the human influence on model building, data collection, and on the outcome variable when it pertains to humans.
Keywords: behavioral science, prediction, feedback model

99. Doran, J.C. & Horst, T.W. (1985). An evaluation of Gaussian plume-depletion models with dual- tracer field measurements. *Atmospheric Environment* **19**, 939-951.
Describes a field measurement program for the evaluation of plume depletion models using simultaneously released, depositing and nondepositing tracers. The models were evaluated according to the ability to predict the ratio of the depositing to nondepositing tracer concentrations.
Keywords: air pollution, structure, input, sensitivity, prediction, prediction measures, graphics, tracer

100. Doss, H. (1989). On estimating the dependence between two point processes. *Annals of Statistics* **17**, 749-763.
Mathematical treatment of the problems of testing for independence between two stationary bivariate point processes and estimating the strength of dependence.
Keywords: statistics, time series, spatial model, cross-correlation, stationary

101. Downing, D.J., Gardner, R.H. & Hoffman, F.O. (1985). An examination of response--surface methodologies for uncertainty analysis in assessment models. *Technometrics* **27**, 151-163.
Discusses approximate sensitivity analyses, where the true model is replaced with a response surface. Sensitivity is determined either by mathematical error propagation analysis or by computer experiments using Latin hypercube sampling of the parameter space. Part of the paper is criticized by R. Easterling (1986, *Technometrics*, 28:91-2).
Keywords: statistics, structure, output, sensitivity, empirical, mechanistic, computer experiment, uncertainty

102. Downton, M.W. & Dennis, R.L. (1985). Evaluation of urban air quality models for regulatory use: refinement of an approach. *Journal of Climate and Applied Meteorology* **24**, 161-173.
Discusses model assessment when use of the model is confined to a few high-ozone days and is focused on prediction of a daily maximum ozone concentration, where the model must be able to replicate concentrations under meteorological conditions specific to a given day and able to reproduce the entire hour-by-hour diurnal pattern of ozone concentrations on a given day. Excellent general overview.
Keywords: output, performance analysis, serial correlation, spatial correlation, sensitivity, graphics, prediction, air pollution, extreme events, review

103. Draxler, R.R. (1987). Accuracy of various diffusion and stability schemes over Washington, D.C. *Atmospheric Environment* **21**, 491-499.
Uses tracer experiment data for evaluating air pollution models. Various graphs and measures are used. Potential temporal and spatial correlations not mentioned.
Keywords: output, prediction, prediction measures, graphics, air pollution, tracer

104. Dubes, R. & Jain, A.K. (1979). Validity studies in clustering methodologies. *Pattern Recognition* **11**, 235-254.
Comprehensive review. Cites 90 articles in 30+ journals.
Keywords: structure, statistics, cluster analysis, inherent error, review, bibliography

105. Duffy, D.E. (1990). On continuity-corrected residuals in logistic regression. *Biometrika* **77**, 287-293.
Presents evidence against the use of continuity corrections for residuals in logistic regression. Compares various residuals with respect to both their null behavior when the fitted model is true and their ability to detect contamination.
Keywords: statistics, logistic regression, diagnostics, graphics, structure, residual analysis

106. Duong, Q.P. (1989). The combination of forecasts: a ranking and subset selection approach. *Mathematical and Computer Modelling* **12**, 1131-1143.
Suggests methods for screening out bad forecast models using subset selection methods from the ranking and selection literature. The remaining forecasts are combined.
Keywords: time series, ranking and selection, forecasting, prediction, selection, prediction measures, testing

107. Dyksen, W.R., Houstis, E.N., Lynch, R.E. & Rice, J.R. (1984). The performance of the collocation and Galerkin method with Hermite bi-cubics. *SIAM Journal on Numerical Analysis* **21**, 695-715.
Uses statistical methods to evaluate the performances of some specific numerical algorithms for solving partial differential equations.
Keywords: structure, numerical analysis, statistics, performance analysis, mathematics, mechanistic

108. Edmonston, B. (1987). Stability and validity of cluster dissection of U.S. life insurance companies. *American Statistical Association Proceedings of the Social Statistics Section* 465-470.
Discusses criteria for assessing the validity of a classification scheme. Applies jackknife and bootstrap procedures. The criteria were applied to a cluster analysis of insurance companies based on four variables.
Keywords: statistics, structure, cluster analysis, bootstrap, jackknife, economics

109. Efron, B. (1986). How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association* **81**, 461-470.
Discusses the downward bias of apparent error in regression analysis when apparent error is estimated from the training data. Gives formulas for estimating the bias. Presents a parametric bootstrap method for estimating the bias in the logistic and normal least squares cases.
Keywords: statistics, empirical, prediction, bootstrap, cross-validation, prediction measures, diagnostics, selection, uncertainty, regression, logistic regression

110. Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy with comments and reply. *Statistical Science* **1**, 54-77.
Excellent review of bootstrap procedures for calculating standard errors and confidence intervals. Discusses relationship to the jackknife and delta methods. Many examples.
Keywords: statistics, bootstrap, uncertainty, review

111. Ehrendorfer, M. & Murphy, A.H. (1988). Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Monthly Weather Review* **116**, 1757-1770.
Discusses the concept of sufficiency, due to DeGroot and Fienberg, and illustrates its interpretations and implications in the context of comparative evaluation of weather forecasting systems. Presents a useful graphical description of sufficiency.
Keywords: meteorology, forecasting, prediction measures, statistics, graphics, output, prediction

112. Einhorn, H.J. & Hogarth, R.M. (1982). Prediction, diagnosis, and causal thinking in forecasting. *Journal of Forecasting* **1**, 23-36.
provides interpretations of measures of forecast accuracy, interpretations that depend on how one's actions affect outcomes and whether one realizes the extent of those effects. Presents a structure for building a causal model for the process.
Keywords: input, selection, prediction, philosophy, causal effects, forecasting, feedback model

113. Eubank, R.L. & Spiegelman, C.H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association* **85**, 387-393.
To test the appropriateness of a simple linear regression model, the authors propose a test based on a cubic spline nonparametric fit. The critical value is determined by Monte Carlo or bootstrapping techniques.
Keywords: statistics, regression, goodness-of-fit, diagnostics, structure

114. Feldman, R.M., Curry, G.L. & Wehrly, T.E. (1984). Statistical procedure for validating a simple population model. *Environmental Entomology* **13**, 1446-1451.
Describes a test that the mean of the vector of differences is zero when the elements of the vector are correlated. Discusses strategy if covariances are unknown.
Keywords: output, prediction, ecology, mechanistic, statistics, serial correlation

115. Fikus, F. & Pawlowski, P. (1989). Bioelectrorheological model of the cell. 2: Analysis of creep and its experimental verification. *Journal of Theoretical Biology* **137**, 365-373.
The model is evaluated by plotting predicted and observed versus time.
Keywords: theoretical biology, dynamic, deterministic, graphics, output

116. Finkel, A.M. (1990). Confronting uncertainty in risk management, A report from the Center for Risk Management, Resources for the Future, 1616P Street, N.W., Washington, D.C. 20036. (UnPub)
Describes uncertainties in risk assessment. Recommends Lorenz curves. Concludes that one should not average the results of incompatible theories, except in special circumstances.
Keywords: uncertainty, output, ecology, biomedical

117. Finkelstein, L. & Carson, E.R. (1985). *Mathematical modelling of dynamic biological systems*, 2nd ed. New York:John Wiley & Sons, Inc..
Presents general principles for formulating mathematical models for linear time invariant dynamic systems with a single input and a single output (compartment models). Uses signal/noise/output terminology when discussing model testing.
Keywords: theoretical biology, mechanistic, structure, dynamic, testing, general

118. Fisher, N.I. (1983). Graphical methods in nonparametric statistics: a review and annotated bibliography. *International Statistical Review* **51**, 25-58.
Reviews the range of graphical methods available for use with nonparametric procedures. Includes an annotated bibliography.
Keywords: statistics, testing, output, prediction, graphics, bibliography, review

119. Florens, J.P. & et, a.l. (1983). *Specifying Statistical Models from Parametric to Non-parametric, Using Bayesian or Non-Bayesian Approaches [Vol. 16 in Lecture Notes in Statistics]*, New York:Springer-Verlag.
Contains 12 mathematical statistics papers that were delivered at a conference in Europe. Topics include model specification, measures of distance between models, robustness, diagnostics, and adaptivity. Half the papers take the Bayesian perspective.
Keywords: statistics, empirical, structure, Bayesian, specification, selection

120. Fowlkes, E.B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74**, 503-515.
Describes diagnostic methods, quantitative and graphical, for use with binary logistic regression.
Keywords: structure, diagnostics, graphics, empirical, statistics, logistic regression

121. Fox, D.G. (1981). Judging air quality model performance. *Bulletin of the American Meteorological Society* **62**, 599-609.
Summary of the American Meteorological Society Workshop on Dispersion Model Performance, Woods Hole, 8-11 Sept., 1980 - sponsored by the EPA. Contains an appendix of notation and terminology. Excellent review of modeling, uses of models, and assessment of models.
Keywords: input, output, prediction measures, statistics, meteorology, review, general

122. Fox, D.G. (1984). Uncertainty in air quality modeling. *Bulletin of the American Meteorological Society* **65**, 27-36.
Summarizes the AMS/EPA Workshop, Woods Hole, MA, Sept., 1982, held to review uncertainty in air quality modelling. Contains a nice mathematical discussion of uncertainty. Identifies components of reducible uncertainty. Discusses problem when both the observed and predicted values are zero. Concludes with a list of recommendations.
Keywords: meteorology, inherent error, serial correlation, bootstrap, review, uncertainty

123. Freedman, D.A. (1985). Statistics and the scientific method. In *Cohort Analysis in Social Research*, Eds. Mason, W.M. & Fienberg, S.E., pp. 343-390. New York: Springer-Verlag.
Criticizes most statistical work in the social sciences because the underlying stochastic models have not been validated. Explains that models are important as data descriptive devices. With discussion.
Keywords: regression, behavioral science, empirical, mechanistic, goodness-of-fit

124. Freedman, L.S. & Pee, D. (1989). Return to a note on screening regression equations. *American Statistician* **43**, 279-282.
Shows that a large ratio of predictors to cases makes it difficult to screen out only the unimportant predictor variables. Selection strategies are influenced by correlation among the predictors.
Keywords: statistics, empirical, stepwise regression, selection, structure

125. Gan, F.F. & Koehler, K.J. (1990). Goodness-of-fit tests based on p-p probability plots. *Technometrics* **32**, 289-303.
Discusses p-p plots and tests based on the linearity of those plots for assessing the accuracy of proposed probability models.
Keywords: statistics, goodness-of-fit, stochastic, prediction measures, output, testing

126. Gass, S.I.(Ed.) (1980). *Validation and Assessment Issues of Energy Models*, Washington, D.C.:National Bureau of Standards.
Lady presents an overview of model validation and provides a list of tasks that make up a model validation project. He notes that the literature reflects the immature nature of assessment practices by failing to display a uniform terminology. House and Ball present a generally negative discussion of the goals, strategies, and feasibility of model validation in the context of large energy models developed for policy making. With discussion. Kresge describes an M.I.T. group's approach to assessment of large scale energy models step-by-step. Gass reports the opinions of a specially commissioned panel of 39 experts. In the view of the panel, the following five areas deserve strong support from research funding agencies: (i) Data collection and availability for model development. (ii) Model documentation plan and guidelines. (iii) Model verification and validation plan. (iv) Relationship between model user and developer. (v) Modeling forums of users and developers. Greenberg and Murphy, using an application of Arrow's Possibility Theorem, shows that one does not have the option of developing an ordinal approach to aggregating component comparisons of goodness. Stated, "model assessment costs about as much as model development." Weyant lists the steps in model assessment and forum analysis and identifies those activities that should be done by a third party audit. Greenberger presents two ways of organizing "model analysis" activities, (i) a flow chart of steps in model assessment (evaluation of a single model's performance over a variety of scenarios) and (ii) a forum analysis (comparative evaluation of many models as applied to a single scenario). He mentions the emergence of a field of model analysis, comprising third-party analysts. Shaw describes the documentation required for evaluation of models and the features required for a model to be accessible to potential users. Estimates that documentation takes 25% of the resources that are consumed in model development and testing. Mayer gives a thoughtful overview of the components of modelling, model validation, and model uses. Greenberg suggests a scheme designed to lead to an efficient retrieval scheme that would be of use in model validation activities. Only a very general overview is presented.
Keywords: energy, expert opinion, general, mathematics, structure, mechanistic, empirical, input, output, uncertainty, documentation, access, third party, prediction measures, forecasting, prediction

127. Gass, S.I. (1983). Decision-aiding models: Validation, assessment, and releated issues for policy analysis. *Operations Research* 31, 603-631.
Reviews model building and model validation terminology as used in operations research. Provides a check list of assessment issues. Discusses role of a third party assessor.
Keywords: management, engineering, review, input, output, structure, general

128. Gauch, H.G., Jr. (1988). Model selection and validation for yield trials with interaction. *Biometrics* **44**, 705-715.
Describes alternative methods for modelling the interaction terms in a two-way analysis of variance. Hold-out data and cross-validation techniques are used to evaluate the models.
Keywords: statistics, structure, regression, hold-out data, cross-validation, empirical

129. Gibbons, J.D., Olkin, I. & Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology*, New York:John Wiley and Sons.
Describes indifference zone methods and subset selection methods for a variety of standard situations. In addition to theory, examples are describes and tables for doing the calculations are provided.
Keywords: statistics, ranking and selection, prediction measures, testing

130. Giere, R.N. (1988). *Explaining Science: A Cognitive Approach*, Chicago:University of Chicago Press.
Reviews alternative theories of science, then discusses models & theories, scientific judgment, and models & experiments. Views scientific advancement as evolutionary.
Keywords: philosophy, empirical, mechanistic, Bayesian, testing

131. Godfrey, L.G. & Tremayne, A.R. (1988). Checks of model adequacy for univariate time series models and their application to econometric relationships. *Econometric Reviews* **7**, 1-42.
Reviews tests for the adequacy of autoregressive, moving average models in econometrics. With discussion. 100+ references.
Keywords: economics, testing, misspecification, time series, serial correlation, structure, diagnostics, review, bibliography

132. Goel, P.K. & Zellner, A. (1986). *Bayesian Inference and Decision Techniques*, New York:North-Holland.
Berger blends the best of Bayesian and frequentist methods into a practical approach. A key feature is a sensitivity analysis of the influence of the choice of prior. Clayton et al. conclude that, if the true, or approximately true, model is included among the alternatives considered, all reasonable model selection procedures will possess similar predictive capabilities. Winkler argues that good probability appraisers should be coherent, calibrated, and expert; and he discusses scoring rules. DeGroot and Fienberg provide a mathematical investigation of measures appropriate for evaluating forecasters or forecast systems. They determine that a forecaster should be calibrated and refined.
Keywords: Bayesian, prediction measures, prediction, statistics, selection, structure, forecasting, expert opinion, general, sensitivity, uncertainty

133. Gower, J.C. (1988). Statistics and agriculture. *Journal of the Royal Statistical Society, Series B* **151**, 179-200.
Historical review of statistics in agricultural sciences. A section on modelling is included.
Keywords: agriculture, statistics, historical, general, review

134. Graham, N.E. & Michaelsen, J. (1987). An investigation of the El Nino-southern oscillation cycle with statistical models 2. Model results. *Journal of Geophysical Research* **92**, 14271-14289.
Fits surface sea temperature models. Makes extensive use of cross-validation techniques.
Keywords: earth science, structure, empirical, serial correlation, cross-validation

135. Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health* **79**, 340-349.
Critical evaluation of empirical modeling of observational data as used in epidemiology. Discusses variable selection, control of confounders, interaction, etc. Suggests a change-in-variable method in place of stepwise regression.
Keywords: biomedical, structure, empirical, statistics, diagnostics, selection, stepwise regression

136. Greenstein, J.S. & Revesman, M.E. (1986). Development and validation of a mathematical model of human decisionmaking for human-computer communication. *IEEE Transactions on Systems,Man,and Cybernetics* **16**, 148-154.
Attempt to predict human actions in a dynamic multiple-task decisionmaking situation. The model is based on a discriminant analysis optimization and is evaluated using volunteer college students.
Keywords: prediction, output, classification, mechanistic, empirical, software, control, behavioral science

137. Grigalunas, T.A., Opaluch, J.J., French, D.P. & Reed, M. (1989). Perspective on validating the natural resource damage assessment model system. *Oil and Chemical Pollution* **5**, 217-238.
Compare predictions from the Natural Resource Damage Assessment Model for Coastal and Marine Environments (NRDAM/CME) to data gathered in four oil spills. Concludes that the model predicted to within an order of magnitude accuracy.
Keywords: ecology, data quality, environmental, uncertainty

138. Gupta, S.S. & Huang, D. (1988). Selecting important independent variables in linear regression models. *Journal of statistical planning and inference* **20**, 155-167.
Proposes an automatic variable selection procedure based on ranking and selection principles and a criterion similar to Mallow's Cp.
Keywords: statistics, selection, structure, regression, ranking and selection

139. Gupta, S.S. & Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*, New York:John Wiley and Sons.
Collection of statistical methods for ranking and selection. The presentation is very mathematical, showing formulas and tables, but no numerical examples.
Keywords: statistics, testing, selection, prediction measures

140. Haan, C.T. (1982). *Hydrologic Modeling of Small Watersheds*, St. Joseph, MI:American Society of Agricultural Engineers.
Discusses a variety of specific mathematical models.
Keywords: hydrology, general

141. Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics* **14**, 1453-1462.
Theoretical demonstration that the coverage of percentile- t bootstrap confidence intervals is not improved by taking a large number of bootstrap samples.
Keywords: statistics, bootstrap, uncertainty

142. Hall, P. & Titterington, D.M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *Journal of the Royal Statistical Society, Series B* **51**, 459-467.
Argues that Monte Carlo testing is a viable, maybe even preferable, alternative to asymptotic methods.
Keywords: statistics, output, testing, simulation, bootstrap, permutation test

143. Hall, S.G. (1986). Estimating the uncertainty of the simulation properties of large nonlinear econometric models. *Applied Economics* **18**, 985-993.
Computer simulations to assess the uncertainties in forecasting when parameters are altered in a random manner dependent on the uncertainties in calibration. Some parameter settings were chosen to describe various policy decision.
Keywords: economics, forecasting, simulation, calibration, uncertainty, sensitivity

144. Hand, D.J. (1986). Recent advances in error rate estimation. *Pattern Recognition Letters* **4**, 335-346.
A review and bibliography of methods for estimating the rate of misclassification. Parametric and nonparametric methods are described. Discusses smoothing or average conditional error rates, the potential benefits of taking a weighted average of error rate estimates, and bounds on error rates. 60+ references.
Keywords: output, classification, statistics, bootstrap, cross-validation, jackknife, review, bibliography

145. Hanna, S.R., Egan, B.A., Vaudo, C.J. & Curreri, A.J. (1984). A complex terrain dispersion model for regulatory applications at the Westvaco Luke mill. *Atmospheric Environment* **18**, 685-699.
A model was used to predict hour averaged SO2 concentrations at 11 monitoring stations. Only 10% of the model evaluation data had been used for model development and calibration. Only block 3-hour and 24-hour average concentrations were studied. The highest-second-highest predicted and observed concentrations, unpaired in time, were emphasized in the model evaluations.
Keywords: air pollution, prediction, prediction measures, graphics, mechanistic, extreme events

146. Harnack, R., Kluepfel, C. & Livezey, R. (1986). Prediction of monthly 700 mb heights using simulated medium- range numerical forecasts. *Monthly Weather Review* **114**, 1466-1480.
Using a time series of observations available at grid points, assesses the skill with which forecasts can be made at individual grid points. Discusses problems of artificial skill due to the stepwise selection of the regression model and due to spatial and temporal correlations.
Keywords: meteorology, serial correlation, prediction, prediction measures, spatial correlation, stepwise regression, empirical

147. Harrison, A.J. & Cairnie, L.R. (1988). The development and experimental validation of a mathematical model for predicting hot-surface autoignition hazards using complex chemistry. *Combustion and Flame* **71**, 1-21.
Describes a model related to the likelihood of an accident. Model output was based on a numerical model. Parameters are taken from the literature and subjective ranges of uncertainty are attached. Includes sensitivity analysis and evaluation against data.
Keywords: structure, output, sensitivity, prediction, mechanistic, engineering, graphics, numerical analysis, uncertainty

148. Hashimoto, T., Stedinger, J.R. & Loucks, D.P. (1982). Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. *Water Resources Research* **18**, 14-20.
For the case of stochastic output, the authors propose output summary measures called reliability, resiliency, and vulnerability.
Keywords: water research, output, prediction measures

149. Hawkins, N.C. & Graham, J.D. (1988). Expert scientific judgment and cancer risk assessment: A pilot study of pharmacokinetic data. *Risk Analysis* **8**, 615-625.
Describes an objective approach to selection of experts on a specific issue.
Keywords: biomedical, expert opinion, uncertainty

150. Hayes, S.R. & Moore, G.E. (1986). Air quality performance: a comparative analysis of 15 model evaluation studies. *Atmospheric Environment* **20**, 1897-1911.
Checks the effects of aggregating data in different ways, especially the length of time used in time averaging. Investigates the variation in spatial correlations as the time for averaging is changed. Discusses the inherent limits to model accuracy.
Keywords: prediction, output, review, extreme events, air pollution, inherent error, uncertainty

151. Henrion, M. & Fischoff, B. (1986). Assessing uncertainty in physical constants. *American Journal of Physics* **54**, 791-797.
This informative paper discusses the extent to which assessments of uncertainty have been overly optimistic. Suggests a scheme for improving uncertainty assessment.
Keywords: physics, uncertainty

152. Hernandez, D.B. (1981). On some guiding principles in mathematical modelling with special emphasis on determinism. *Mathematical Modelling* **2**, 179-190.
Describes modeling and model validation from an abstract, mathematical perspective. Mathematical terms, such as initial condition, boundary condition, well-posed, and dynamical system, are defined. The author discusses, in a very general way, how one might validate deterministic and stochastic models.
Keywords: mathematics, general, deterministic, causal effects, dynamic

153. Hill, B.M. (1986). Some subjective Bayesian considerations in the selection of models with comments and reply. *Econometric Reviews* **4**, 191-288.
Presents arguments for the Bayesian approach to modelling. Acknowledges that a prior may be incorrect and suggests revision of the prior after seeing data. Discussion follows the paper. The importance of doing sensitivity analyses is noted.
Keywords: predictive distributions, prediction, output, Bayesian, statistics, selection, empirical, economics

154. Hill, J.R. (1990). A general framework for model-based statistics. *Biometrika* **77**, 115-126.
This mathematical statistics paper suggests a comprehensive theory for data analysis, model building, and model evaluation in which statisticians are encouraged to use both Bayesian and frequentist ideas.
Keywords: statistics, Bayesian

155. Hinkley, D.V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society, Series B* **50**, 321-337.
Survey of developments in bootstrap methodology. Paired with DiCiccio and Romano (1988) and with discussion of both papers.
Keywords: statistics, bootstrap, uncertainty, testing, jackknife, regression, time series, output

156. Hirschberg, J.G. (1987). The relationship between various criteria for the evaluation of flexible functional forms. *American Statistical Association Proceedings of the Business and Economic Statistics Section* 43-48.
Proposes three measures for evaluating econometric models, two based on information measures and one based on the log likelihood. Discusses an example where it is not feasible to use out of sample criteria, such as hold-out data or cross-validation.
Keywords: economics, hold-out data, prediction measures

157. Hirtzel, C.S., Corotis, R.B. & Quon, J.E. (1982). Estimating the maximum value of autocorrelated air quality measurements. *Atmospheric Environment* **16**, 2603-2608.
Determines the effective sample size for calculating the probability that the maximum value in a set of autocorrelated observations exceeds a predetermined standard, assuming that adjacent (in time) pairs of observations are bivariate normal and that the process follows a special "pseudo-Markov" property.
Keywords: meteorology, output, prediction measures, statistics, serial correlation, extreme events

158. Hjorth, U. (1989). On model selection in the computer age. *Journal of statistical planning and inference* **23**, 101-115.
A general mathematical structure and terminology are suggested for the model selection procedure when a finite number of specified models are being considered.
Keywords: statistics, general, selection

159. Hodges, J.S. (1987). Uncertainty, policy analysis and statistics. *Statistical Science* **2**, 259-291.
Excellent overview of model building, uncertainties in the model, and how those uncertainties affect conclusions based on the model. Discussion follows the paper.
Keywords:, input, structure, output, numerical analysis, prediction, statistics, mechanistic, empirical, predictive distributions, bootstrap, Bayesian, uncertainty

160. Holland, J.Z. (1988). Truncation effects on estimated parameters of tracer distributions sampled on finite domains. *Journal of Applied Meteorology* **27**, 280-294.
Reviews parametric and nonparametric methods for estimating the cross-wind distribution and the mean and standard deviation of the distribution based on tracer releases at a single source.
Keywords: meteorology, tracer, structure, substructure, prediction, empirical

161. Horowitz, J. (1980). Extreme values from a nonstationary stochastic process: an application to air quality analysis. *Technometrics* **22**, 469-482.
For the model $\log x(t) = f(t) + e(t)$, where $f(t)$ is deterministic and the stochastic $e(t)$ is stationary and normal, the author derives the extreme value distribution. Simulations were used to check and illustrate the theory.
Keywords: air pollution, extreme events, nonstationary, simulation, statistics

162. Hunter, W.G. & Jaworski, A.P. (1986). A useful method for model-building II: synthesizing response functions from individual components. *Technometrics* **28**, 321-327.
In some cases, the response variable y is not directly measured but is created from the measured responses s,t,u,v,..., where y=g(s,t,u,v,...). These authors suggest that building models separately for the component responses, s,t,u,v,.., and then taking the g(.) function of those component models to create the surrogate y. Examples are given.
Keywords: structure, statistics, regression, substructure, empirical, mechanistic, output, surrogate

163. Iman, R.L. & Davenport, J.M. (1982). Rank correlation plots for use with correlated input variables. *Communications in Statistics,Part B-Simulation and Computation* **11**, 335-360.
Demonstrates a procedure for generating multivariate input data having fixed marginal distributions and specified rank correlation coefficients.
Keywords: input, sensitivity, computer experiment

164. Iman, R.L., Helton, J.C. & Campbell, J.E. (1981). An approach to sensitivity analysis of computer models: part II--ranking of input variables, response surface validation, distribution effect and technique synopsis. *Journal of Quality Technology* **13**, 232-240.
Proposes approximating a complicated, mechanistic model with a response surface based on polynomials and then performing sensitivity analyses using the response surface. This imposes an empirical model building and validation phase to find an acceptable response surface equation. Provides a good overview of this computer experimentation approach to studying large, mechanistic models.
Keywords: statistics, sensitivity, structure, graphics, computer experiment, diagnostics, regression, empirical, mechanistic,

165. Irwin, J. & Smith, M. (1984). Potentially useful additions to the rural model performance evaluation. *Bulletin of the American Meteorological Society* **65**, 559-568.
A comparison of 10 models for predicting SO_2 concentrations. Refines the comparisons by studying individual pollution episodes and segregating the data by stability class.
Keywords: graphics, air pollution, prediction, sensitivity, input, output

166. Irzik, G. & Meyer, E. (1987). Causal modeling: new directions for statistical explanation. *Scandinavian Journal of Statistics* **54**, 495-514.
Discusses various cause-effect modelling techniques, such as path analysis and structural equations, with the goal of determining the extent to which statistical association can imply cause. It is shown that a pattern of path coefficients can arise from a variety of different models and that Simpson's Paradox can lead to invalidate conclusions if an unknown variable is lurking. Concludes that causal modelling methods cannot pretend to deduce causal information from statistics alone; their value lies rather in the means they provide to evaluate and test the implications of models.
Keywords: causal effects, philosophy, statistics, explanation

167. Jacks, E. & Rao, S.T. (1985). An examination of the MOS objective temperature prediction model. *Monthly Weather Review* **113**, 134-148.
Uses a hold-out data set in evaluation of a multiple linear regression technique for predicting temperatures. Calculates a wide variety of statistics and constructs many different graphs, some based on paired and some based on unpaired data.
Keywords: meteorology, output, prediction, prediction measures, graphics, bootstrap, uncertainty, regression, extreme events, hold-out data

168. Jacoby, S.LS. & Kowalik, J.S. (1980). *Mathematical Modeling with Computers*, Englewood Cliffs, NJ:Prentice-Hall, Inc..
Discusses the model building process and defines the terminology used in model-building and validation. Describes potential sources of bias and uncertainty due to converting the conceptual model into mathematical form and then converting that into a computer program.
Keywords: computer model, numerical analysis, mathematics, general, simulation

169. Jakeman, A.J., Simpson, R.W. & Taylor, J.A. (1988). Modeling distributions of air pollutant concentrations- III. The hybrid deterministic-statistical distribution approach. *Atmospheric Environment* **22**, 163-174.
A causal (mechanistic) model is used to predict order statistics of the distribution of pollutant concentrations at individual receptor locations. States that, although typical conditions can often be well-modelled by fundamental laws, extreme conditions can only be estimated statistically. Discusses calibration in the face of doubly censored data.
Keywords: air pollution, output, structure, mechanistic, empirical, calibration, uncertainty, prediction, extreme events

170. Joglekar, G., Schuenemeyer, J.H. & LaRiccia, V. (1989). Lack-of-fit testing when replicates are not available. *American Statistician* **43**, 135-143.
Reviews two methods: (i) nearest neighbor grouping followed by a lack-of-fit F test and (ii) data splitting where one tests for a significant reduction in residual variance when each data subset modeled separately.
Keywords: statistics, goodness-of-fit, regression, testing, selection, structure, prediction measures

171. Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: heuristics and biases*, New York:Cambridge University Press.
Excellent collection of articles concerning human perceptions and intuitions about statistical and empirical issues. Particularly important when assessing the validity of expert opinion, determining causal mechanisms, and communicating the model and model results.
Keywords: statistics, behavioral science, regression, prediction, uncertainty, Bayesian, expert opinion

172. Kass, R.E., Tierney, L. & Kadane, J.B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663-674.
Studies perturbation in the posterior density caused by deleting a case or altering the prior.
Keywords: statistics, diagnostics, Bayesian, sensitivity measures

173. Kelley, J.G., Russo, J.M., Eyton, J.R. & Carlson, T.N. (1988). Mesoscale forecasts generated from operational numerical weather-prediction model output. *Bulletin of the American Meteorological Society* **69**, 7-15.
A nice review of methods for forecasting the weather. An empirical response surface technique is evaluated by comparing forecasts to surface observations. Standard statistical assessment measures are used.
Keywords: meteorology, prediction, prediction measures, graphics, empirical

174. Kerr, R.A. (1989). Hansen vs. the world on the greenhouse threat. *Science* **244**, 1041-1043.
Report on the modelling efforts to predict global climatic changes under different scenarios, particularly those that could lead to global warming. The article identifies different views on the uncertainties attached to the model output and it points out the difficulties of validating these models.
Keywords: ecology, prediction, uncertainty, output

175. Khamis, H.L. (1990). The δ-corrected Kolmogorov-Smirnov test for goodness of fit. *Journal of statistical planning and inference* **24**, 317-335.
Proposes alternative versions of empirical distribution function plots for purposes of calculating the Kolmogorov-Smirnov statistic. Includes is a review of methods for plotting the e.d.f.
Keywords: statistics, goodness-of-fit, historical, testing, output, prediction measures, graphics

176. Khan, M.H. (1989). Evaluation of rainfall-discharge models with discrimination. *Journal of Hydrology* **108**, 63-78.
Two calibrations were compared and the models are evaluated against independent data.
Keywords: hydrology, calibration, prediction, prediction measures, output

177. Khuri, A.I. (1986). Exact tests for the comparison of correlated response models with an unknown dispersion matrix. *Technometrics* **28**, 347-357.
Reviews methods for testing the equality of parameters from several correlated linear models, a problem arising in repeated measures experiments.
Keywords: statistics, testing, spatial correlation, serial correlation, output

178. King, D.S. & Bunker, S.S. (1984). Application of atmospheric models for complex terrain. *Journal of Climate and Applied Meteorology* **23**, 239-246.
Evaluates two atmospheric transport models using tracer experiments and paired observed and predicted values.
Keywords: air pollution, tracer, prediction, graphics, prediction measures

179. Kmenta, J. & Ramsey, J.B.(Eds.) (1980). *Evaluation of econometric models*, New York, New York:Academic Press.
The introductory chapter by Ramsey and Kmenta covers the main issues in model evaluation. Chow reviews econometric modelling, especially the problem of choosing a loss function for a feedback model when the purpose is to achieve optimum control. Kadane and Dickey discuss their belief that utility functions and Bayesian reasoning provide the best approach to model evaluation in economics.
Keywords: economics, output, structure, control, dynamic, feedback model, specification, time series, prediction, Bayesian, selection, regression

180. Knopman, D.S. & Voss, C.I. (1988). Discrimination among one-dimensional models of solute transport in porous media: Implications for sampling design. *Water Resources Research* **24**, 1859-1876.
Proposes an iterative process for comparing models. Criteria are a battery of goodness-of-fit measures and residual analyses. The design for gathering new data is chosen to maximize the differences between models.
Keywords: water research, design, mechanistic, structure, selection, spatial correlation, prediction, prediction measures, diagnostics

181. Kokkelenberg, E.C. & Nguyen, S.V. (1987). Forecasting comparison of three flexible functional cost forms. *American Statistical Association Proceedings of the Business and Economic Statistics Section* 57-65.
Provides a procedure, based on hold-out data, for comparing models. Checks the substructures for face validity.
Keywords: forecasting, economics, prediction, prediction measures, face validity, hold-out data

182. Kramer, W. & Sonnberger, H. (1986). *The Linear Regression Model Under Test*, Heidelberg:Physica-Verlag.
Complete discussion of statistical diagnostic and selection procedures for multiple regression models, especially when there is a time ordering of the observations. Cites 200+ references.
Keywords: regression, testing, statistics, specification, structure, time series, selection, economics, diagnostics

183. Kramer, W., Sonnberger, H., Maurer, J. & Havlik, P. (1985). Diagnostic checking in practice. *The Review of Economics and Statistics* **67**, 118-123.
Some empirical models are submitted to a battery of tests.
Keywords: economics, statistics, testing, prediction measures, prediction, time series, diagnostics, empirical

184. Ku, J., Rao, S.T. & Rao, K.S. (1987). Numerical simulation of air pollution in urban areas: model performance. *Atmospheric Environment* **21**, 213-232.
Evaluates model performance using a variety of graphs and measures of prediction and using both paired and unpaired predicted and observed values. No mention is made of serial or spatial correlations.
Keywords: prediction measures, prediction, output, bootstrap, graphics, air pollution, testing, uncertainty, dynamic

185. Lam, N.S. (1981). The reliability problem of spatial interpolation models. In *Modeling and Simulation, Volume 12*, Eds. Vogt, W.G. & Mickle, M.H., pp. 869-876. Research Triangle Park, NC: The Instrument Society of America.
Generates a range of fractal surfaces of varying complexity and then applies distance--weighting and trend surface analyses to see which works better. Distance-weighting and stratified systematic unaligned sampling were superior techniques.
Keywords: spatial model, empirical, structure, prediction, prediction measures, statistics, earth science

186. Lam, N.S. (1983). Spatial interpolation methods: a review. *American Cartographer* **10**, 129-149.
Presents a comprehensive review of approaches to the spatial interpolation problem of finding the function that will best represent the whole surface and that will predict values at points or for subareas.
Keywords: spatial correlation, spatial model, empirical, review, structure, historical

187. Landry, M., Malouin, J.L. & Oral, M. (1983). Model validation in operations research. *European Journal of Operational Research* **14**, 207-220.
General description of model validation with particular reference to models as used for energy policy formulation. Discusses model confidence, which is an expression of the user's total attitude toward the model, and of the willingness to employ its results in making decisions.
Keywords:, general, engineering, energy, input, structure, output, philosophy

188. Lanzante, J.R. (1984). Strategies for assessing skill and significance of screening regression models with emphasis on Monte Carlo techniques. *Journal of Climate and Applied Meteorology* **23**, 1454-1458.
Discusses artificial predictability, artificial skill, true skill, hindcast skill, and forecast skill. Refers to the "effective number of independent gridpoints," "effective sample size," and "time needed to gain another degree of freedom."
Keywords: meteorology, statistics, regression, serial correlation, selection, testing, empirical, spatial correlation

189. Law, A.M. & Kelton, W.D. (1982). *Simulation modeling and analysis*, New York:-McGraw-Hill Company.
In Chapter 10 - Validation of Simulation Models, the authors provide a general discussion of model validation
Keywords: engineering, testing, serial correlation, stationary, output, face validity, software

190. Lawrence, M.J., Edmundson, R.H. & O'Conner, M.J. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science* **32**, 1521-1532.
Empirically examines whether averaging forecasts improves on the accuracy of a single method. Mean absolute percentage error was the chosen measure of accuracy.
Keywords: empirical, prediction, output, prediction measures, forecasting

191. Le, D.N. & Petkau, A.J. (1988). The variability of rainfall acidity revisited. *Canadian Journal of Statistics* **16**, 15-38.
Describes an attempt to validate a kriging model proposed for creating a contour map based on data independent of that used to form the original model.
Keywords: air pollution, water research, spatial model, time series, input, structure, selection, output, prediction, prediction measures, graphics, empirical

192. Leamer, E.E. (1978). *Specification Searches*, New York:John Wiley & Sons, Inc..
Presents Bayes and empirical Bayes methodology for use with observational data. The phrase "specification searching" is used for the complex process by which nonexperimental scientists choose a model. Hypothesis testing and multiple regression receive extensive discussion. Thoughtful, instructive publication.
Keywords: economics, behavioral science, statistics, structure, Bayesian, empirical, sensitivity, regression, empirical Bayes, selection, prediction, control, specification, testing

193. Lee, I.Y. (1989). Evaluation of cloud microphysics parameterizations for mesoscale simulations. *Atmospheric Research* **24**, 209-220.
Reviews published rate parameters and evaluates them for conformation to the theoretical principles of cloud and precipitation physics and for their applicability to regional models. Generates data from a plausible model and expresses the rate parameters as functions of regionally measurable variables, the functions being fit by multiple regression. This is an approach not mentioned in other references.
Keywords: meteorology, calibration, input, regression, surrogate

194. Lehman, R.S. (1977). *Computer Simulation and Modeling: An Introduction*, New York:Lawrence Erlbaum Associates.
Identifies major areas with which validation must be concerned and gives a detailed review of validation procedures.
Keywords: computer model, structure, prediction, simulation, expert opinion, testing, prediction measures

195. Lehmann, E.L. (1990). Model specification: the views of Fisher and Neyman and later developments. *Statistical Science* **5**, 160-168.
Excellent historical and philosophical discussion of the contributions of statistical theory to model specification or construction, especially in the areas of (i) a reservoir of models, (ii) model selection, (iii) and classification of models.
Keywords: statistics, philosophy, historical, general, empirical, mechanistic, selection, stochastic, deterministic,

196. Leimkuhler, K. (1982). Some methodological problems in energy modelling. In *Progress in modelling and simulation*, Ed. Cellier, F.E., pp. 61-73. London, England: Academic Press.
This is a comprehensive review of mathematical modelling as applied to energy and environmental issues. It includes an historical perspective.
Keywords: simulation, structure, prediction, energy, historical, surveillance, control, review

197. Lemeshow, S., Teres, D., Avrunin, J.S. & Pastides, H. (1988). Predicting the outcome of intensive care unit patients. *Journal of the American Statistical Association* **83**, 348-356.

Uses a training set of 755 intensive care unit patients to construct a logistic multiple regression model for predicting life-death outcomes for intensive care patients. Compares the model predictions to the actual outcomes of 1997 independent, subsequent ICU patients.

Keywords: biomedical, prediction, output, empirical, logistic regression

198. Levy, E. (1987). Implementing statistical quality control in forecasting. *American Statistical Association Proceedings of the Business and Economic Statistics Section* 512-517.

Describes how quality control techniques, such as control charts, could be used to improve the quality of model-based forecasts. The primary difficulty is in determining the correct variability.

Keywords: surveillance, statistics, quality control, forecasting, uncertainty

199. Linhart, H. & Zucchini, W. (1986). *Model Selection*, New York:John Wiley & Sons.,Inc..

Discusses the sense in which "overall discrepancy" is the sum of the discrepancies due to approximation and estimation. Discusses measures of discrepancy between two prediction distributions, particularly Kullback-Leibler discrepancy which leads to maximum likelihood estimators and the Akaike Information Criterion for calibration within the approximating family. Describes approximate methods based on large sample theory, cross- validation, or bootstrapping. Treats a number of empirical modelling strategies, including simple and multiple regression, analysis of variance, analysis of covariance, analysis of proportions, and time series models in both the time and frequency domains.

Keywords: statistics, selection, regression, empirical, bootstrap, serial correlation, cross-validation, specification, structure, time series

200. Liu, M. & Moore, G.E. (1984). On the evaluation of predictions from a Gaussian plume model. *Journal of the Air Pollution Control Association* **34**, 1044-1050.

Compares the predictive accuracy of the mechanistic model to that of a naive model. Evaluation focuses on the ability to predict extreme events such as the highest or 5th highest concentration at a site. Discusses the effects of inherent error.

Keywords: prediction, empirical, mechanistic, prediction measures, air pollution, extreme events, inherent error

201. Livezey, R.E. & Chen, W.Y. (1983). Statistical field significance and its determination by Monte Carlo techniques. *Monthly Weather Review* **111**, 46-59.

Suggests a goodness-of-fit statistic and proposes using Monte Carlo simulation to determine significance in a manner that accounts for spatial and temporal correlations. Gives a formula for the effective time between independent samples.

Keywords: output, prediction, prediction measures, time series, meteorology, empirical, serial correlation, spatial correlation, bootstrap, goodness-of-fit

202. Loehle, C. (1983). Evaluation of theories and calculation tools in ecology. *Ecological Modelling* **19**, 239-247.

Excellent review of models, uses of models, and methods of model assessment.

Keywords: structure, mechanistic, empirical, prediction, ecology, goodness-of-fit, testing

203. Lott, R.A. (1986). Model performance-plume dispersion over elevated terrain. *Atmospheric Environment* **20**, 1547-1554.
Focuses on the frequency distribution of the highest measured and highest predicted concentrations over 1 h and 3 h intervals at 8 stations. The models are to be used to predict extreme events, not to predict specific time/space concentrations.
Keywords: air pollution, prediction, prediction measures, mechanistic, extreme events

204. Lotwick, H.W. & Silverman, B.W. (1982). Methods for analyzing spatial processes of several types of points. *Journal of the Royal Statistical Society, Series B* **44**, 406-413.
Proposes a test for the null hypothesis that two spatial point processes (for outputs of types x and y) are independent, with application to real data.
Keywords: statistics, spatial model, structure, empirical, biomedical

205. Lough, J.M. & Fritts, H.C. (1985). The southern oscillation and tree rings: 1600-1961. *Journal of Climate and Applied Meteorology* **24**, 952-966.
Uses tree ring data to reconstruct climate conditions of the past. An empirical model is fit using sophisticated multivariate statistics. Model assessment uses cross-validation, simulation, and cross-power-spectral analysis. Specific details are omitted.
Keywords: meteorology, empirical, output, time series, serial correlation, hold-out data, cross-correlation

206. MacAleer, M. & Veall, M.R. (1989). How fragile are fragile inferences? A re-evaluation of the deterrent effect of capital punishment. *Review of Economics and Statistics* **71**, 99-106.
Describes a technique in which one determines the limits of the estimated regression coefficient of interest over all possible models. These limits indicate the fragility of inference about the focus variable.
Keywords: behavioral science, uncertainty, structure, testing, bootstrap, regression, sensitivity measures, empirical, selection, specification

207. MacEachren, A.M. & Davidson, J.V. (1987). Sampling and isometric mapping of continuous geographic surfaces. *American Cartographer* **14**, 299-320.
Discusses the uncertainty attached to estimates at points on a continuous geographic surface. Suggests sampling schemes and sample sizes. Gives guidelines for identifying the minimum reasonable resolution of isometric mapping.
Keywords: spatial model, empirical, uncertainty, design, output, prediction measures, graphics, earth science

208. Mahmoud, E. (1984). Accuracy in forecasting: a survey. *Journal of Forecasting* **3**, 139-159.
Investigates the accuracy performance of different forecasting techniques and methods for combining separate forecasts. Concludes that forecasters can improve the quality of their predictions by relying on more than one forecasting method.
Keywords: output, prediction, prediction measures, forecasting,

209. Mason, I. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine* **30**, 291-303.
Discusses the receiver operating characteristic curve (ROC) and indices based on the ROC for assessing probabilistic predictions of a binary event.
Keywords: meteorology, prediction, forecasting, prediction measures, output

210. McFadden, D. (1987). What do microeconometricians really do?. *American Statistical Association Proceedings of the Business and Economic Statistics Section* 402-405.
Proposes that econometric modelers pay more attention to cross-validation techniques, including the bootstrap, and to evaluating models against independent data.
Keywords: economics, cross-validation, bootstrap

211. McGinnis, M.A. (1989). Comparative evaluation of freight transportation choice models. *Transportation Journal* **29**, 36-46.
Compares each model against empirical studies, classifying each study as being consistent or inconsistent with the four models.
Keywords: management, face validity

212. McGregor, J.R. & Babb, J.C. (1989). Serially correlated differences in the paired comparison of time series. *Biometrika* **76**, 735-739.
Investigates the effects on an F test when paired differences follow a stationary Markov process with normal residuals. For the mathematical statistician.
Keywords: statistics, time series, testing, output

213. McNees, S.K. (1986). Forecasting accuracy of alternative techniques: a comparison of U.S. macroeconomic forecasts with comments and reply. *Journal of Business and Economic Statistics* **4**, 5-23.
Empirical forecasting models evaluated against specific series of economic data. Discussants point out that policy decisions based on early data in the series may have affected data later in the series.
Keywords: prediction, prediction measures, output, economics, empirical, feedback model

214. McPherson, G. (1989). The scientists' view of statistics - a neglected area. *Journal of the Royal Statistical Society, Series A* **152**, 221-240.
States that statistical models bring structure to scientific thought and that inferential statistics provides an objective, probability-based comparison of model and data. Presents interesting structure and justification for this thesis, with suggestions. With discussion.
Keywords: statistics, testing, output

215. Michaelsen, J. (1987). Cross-validation in statistical climate forecast models. *Journal of Climate and Applied Meteorology* **26**, 1589-1600.
Describes cross-validation methods for variable selection and assessment of prediction accuracy. Discusses hindcast, forecast, and artificial skills. Calculates the time between effectively independent observations.
Keywords: output, cross-validation, empirical, statistics, regression, outliers, selection, serial correlation

216. Mitchell, T.J. & Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023-1036.
Concerns the selection of subsets of predictor variables in a linear regression model for the prediction of a dependent variable. Predictive error is determined using a Bayesian cross-validation approach.
Keywords: statistics, Bayesian, regression, selection, structure, empirical, cross-validation, energy

217. Moore, M. (1988). Predicting iceberg trajectories. *Chance:New Directions for Statistics and Computing* **1**, 30-36.
Uses autoregressive integrated moving average models and exponential smoothing models to forecast iceberg trajectories. Holds out last 20 observations for model evaluation.
Keywords: output, prediction, time series, statistics, empirical, forecasting, hold-out data

218. Mullerburg, M. (1984). Software validation: a bibliography. In *Software Validation*, Ed. Hausen, H., pp. 335-375. New York: North-Holland.
Bibliography of approximately 300 references, with key words and authors lists and an annotated bibliography on assertion-based testing and debugging.
Keywords: testing, bibliography

219. Murphy, A.H. & Winkler, R.L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association* **79**, 489-500.
Interesting historical review of the efforts to create objective probability forecasts and to attach a measurement of uncertainty to those forecasts in meteorology.
Keywords: meteorology, statistics, forecasting, prediction, output, historical, review

220. Murphy, A.H. & Winkler, R.L. (1987). A general framework for forecast verification. *Monthly Weather Review* **115**, 1330-1338.
Uses the joint distribution of forecasts and observations as the basis for a unified framework for forecast verification. Discusses calibration (accuracy) and refinement of a forecaster.
Keywords: meteorology, forecasting, Bayesian, prediction, output, expert opinion

221. Murthy, D.N.P. & Rodin, E.Y. (1987). A comparative evaluation of books on mathematical modelling. *Mathematical Modelling* **9**, 17-28.
Annotated bibliography of 62 books on mathematical modelling. States that the books give too little attention to parameter estimation, design of experiments, and validation. Very few studies validate the mathematical models and still fewer discuss the evolution of the model and the rationale for changes made during the evolutionary process.
Keywords: review, bibliography

222. National Bureau of Standards (1984). Guideline for Lifecycle Validation, Verification, and Testing of Computer Software. Federal Information Processing Standards Publication 101. (UnPub)
Presents guidelines for five phases of software development and maintenance, the definition and analysis phase, the design phase, the programming and testing phase, the installation phase, and the operations and maintenance phase.
Keywords: software, testing, verification, structure

223. Neelamkavil, F. (1987). *Computer Simulation and Modelling*, New York:John Wiley & Sons, Inc..
In depth description of modeling and model validation, including a review of selected philosophies of science.
Keywords: philosophy, input, structure, output,

224. Nelder, J.A. (1979). Experimental design and statistical evaluation. In *Performance evaluation of numerical software*, Ed. Fosdick, L.D., pp. 309-315. Amsterdam, Holland: North-Holland Publishing Company.
Discusses the principles of statistical experimental design as applied to the testing of numerical software.
Keywords: statistics, design, input, numerical analysis, structure, software

225. Neumann, C.J., Lawrence, M.B. & Caso, E.L. (1977). Monte Carlo significance testing as applied to statistical tropical cyclone prediction models. *Journal of Applied Meteorology* **16**, 1165-1174.
Uses Monte Carlo simulation to determine correct critical values for the F-test of the multiple correlation coefficient when a stepwise regression has been used to select from among a large number of predictors in the presence of temporal and spatial correlations.
Keywords: meteorology, prediction, prediction measures, stepwise regression, empirical, selection, serial correlation, spatial correlation

226. Novotny, T.J. & McDonald, L.L. (1986). Model selection using discriminant analysis. *Journal of Applied Statistics* **13**, 159-165.
Discusses selection from among m models in a discriminant analysis problem. The proposed technique uses goodness-of-fit test statistics and relies on data simulated under each model.
Keywords: statistics, structure, specification, selection, ecology, classification, goodness-of-fit, testing

227. Obert, M., Pfeifer, P. & Sernetz, M. (1990). Microbial growth patterns described by fractal geometry. *Journal of Bacteriology* **172**, 1180-1185.
The fractal geometry growth hypothesis was supported by the data. This work suggests new approaches for toxicity tests and growth modeling.
Keywords: theoretical biology, structure, prediction measures, prediction, explanation, discovery

228. Olnick, M. (1981). Mathematical models in the social and life sciences: A selected bibliography. *Mathematical Modelling* **2**, 237-258.
List of 524 publications, classified by subject area. Most entries are pre-1980 books.
Keywords: biomedical, behavioral science, bibliography,

229. Orlob, G.T.(Ed.) (1983). *Mathematical Modeling of Water Quality: Streams, Lakes, and Reservoirs*, New York:John Wiley and Sons.
Beck reviews the modelling process and discusses calibration, identification, methods for combining theoretical knowledge with empirical knowledge, sensitivity analysis, & the extended Kalman filter method of recursive estimation of parameters (and/or state variables). Orlob provides an historical account of mathematical modelling of water quality and suggests areas requiring research.
Keywords: water research, general, historical, input, structure, output, verification, sensitivity, calibration, sensitivity

230. Pederson, S.P. & Johnson, M.E. (1990). Estimating model discrepancy. *Technometrics* **32**, 305-314.
Proposes an estimate of model discrepancy, based on chi-squared, which incorporates both sample size N and the number of bins. Indicates the magnitude of lack-of-fit and information on the pattern of lack-of-fit.
Keywords: statistics, goodness-of-fit, output, testing, prediction measures

231. Peters, D. & Randles, R.H. (1990). A multivariate signed-rank test for the one-sample location problem. *Journal of the American Statistical Association* **85**, 552-557.
Method appropriate for testing for a difference between matched observed and predicted multivariate output variables.
Keywords: statistics, testing, prediction measures, output

232. Peters, S.C. & Freedman, D.A. (1985). Using the bootstrap to evaluate forecasting equations. *Journal of Forecasting* **4**, 251-262.
Describes a bootstrap technique for estimating the standard error of a forecast where the forecast equation is based on a dynamic linear model having serial correlations. Also discusses bootstrapping for selection. The bootstrap is compared to the delta method. No assessment against observed data.
Keywords: structure, output, bootstrap, forecasting, dynamic, serial correlation, selection, uncertainty

233. Pettit, L.I. (1986). Diagnostics in Bayesian model choice. *The Statistician* **35**, 183-190.
Reviews diagnostic checks on Bayesian models, including checks on the prior, the observations, the likelihood, and the design.
Keywords: statistics, Bayesian, structure, diagnostics, selection

234. Phillips, P.C.B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics* **33**, 311-340.
Describes the asymptotic properties of estimators and test statistics for linear regression analyses when the underlying data are serially correlated. Shows that the t-statistic, for testing that the slope is zero, diverges to infinity as the number of observations increases, even when the slope is actually zero. Similarly disconcerting results hold for point estimates and F-tests.
Keywords: output, prediction, prediction measures, statistics, economics, testing, calibration, serial correlation

235. Picard, R.R. & Berk, K.N. (1990). Data splitting. *American Statistician* **44**, 140-147.
Offers some solutions to the problem of deciding which observations to set aside for model evaluation purposes. Discusses the advantages and disadvantages of data splitting relative to cross-validation and bootstrapping.
Keywords: statistics, hold-out data, prediction measures, cross-validation, regression, review

236. Picard, R.R. & Cook, R.D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association* **79**, 575-583.
Investigates the advantages and disadvantages of some algorithms for data-splitting. Concludes that all aspects of model selection (even those that are cross-validatory) should be confined to analysis of the estimation data and that validation data should be reserved solely for assessment.
Keywords: statistics, regression, empirical, hold-out data, cross-validation, bootstrap,

237. Pool, R. (1989). Is it real, or is it Cray?. *Science* **244**, 1438-1440.
Describes some exciting uses of computer models.
Keywords: computer model, simulation, forecasting, management, performance analysis, prediction, control, discovery

238. Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine* **8**, 431-440.
Provides a mathematical definition of a surrogate variable. Method might be applicable when a surrogate output (or input) variable must be used in place of the 'true' variable.
Keywords: statistics, output, surrogate, testing, biomedical

239. Rao, C.R. (1987). Prediction of future observations in growth curve models. *Statistical Science* **2**, 434-471.
Efficiencies of different methods of prediction are assessed by cross-validation. Bayesian and empirical Bayesian methods are used. The discussants offered some provocative and insightful comments.
Keywords: statistics, growth curves, time series, serial correlation, Bayesian, cross-validation, prediction, review

240. Rao, S.T., Sistla, G., Pagnotti, V., Petersen, W.B., Irwin, J.S. & Turner, D.B. (1985). Evaluation of the performance of RAM with the regional air pollution study data base. *Atmospheric Environment* **19**, 229-245.
Evaluates the RAM model as a predictor of sulphur dioxide concentrations according to the recommendations of Fox (1981, 1984).
Keywords: air pollution, output, prediction, prediction measures, bootstrap, serial correlation, uncertainty

241. Rao, S.T. & Visalli, J.R. (1981). On the comparative assessment of the performance of air quality models. *Journal of the Air Pollution Control Association* **31**, 851-860.
Recommends performance measures and graphical techniques for comparing measured and model predicted concentrations.
Keywords: meteorology, output, prediction measures, prediction, statistics, cross-correlation, graphics

242. Rayner, J.C. & Best, D.J. (1989). *Smooth Tests of Goodness of Fit*, New York, NY:Oxford University Press.
Presents the history, derivations, mathematical details, and examples of smooth tests for goodness-of-fit, which are based on nonparametric density estimation and score statistics. It is suggested that the test results be viewed in combination with nonparametric density plots and Q-Q plots.
Keywords: statistics, goodness-of-fit, testing, output, prediction measures, historical, graphics

243. Rayner, J.C.W. & Best, D.J. (1990). Smooth tests of goodness of fit: an overview. *International Statistical Review* **58**, 9-17.
Describes the history of, and some recent developments in, smooth tests of goodness-of-fit, a class of tests based on nonparametric density estimation.
Keywords: statistics, goodness-of-fit, prediction measures, prediction, historical, review

244. Reckhow, K.H. (1988). A comparison of robust Bayes and classical estimators for regional lake models of fish response to acidification. *Water Resources Research* **24**, 1061-1068.
Tries both maximum likelihood and Bayesian approaches to estimation in logistic regression. Resulting models are compared to independent data.
Keywords: statistics, Bayesian, calibration, logistic regression, empirical, ecology

245. Reckhow, K.H. & Chapra, S.C. (1983). Confirmation of water quality models. *Ecological Modelling* **20**, 113-133.
Outlines some philosophical and statistical issues relevant to the problem of confirmation, and recommends appropriate applications of statistical confirmatory criteria. Reviews the philosophy of science, esp., logical empiricism.
Keywords: philosophy, water research, graphics, historical, hold-out data, serial correlation, cross-correlation, testing, goodness-of-fit,

246. Reckhow, K.H., Clements, J.T. & Dodd, R.C. (1990). Statistical evaluation of mechanistic water-quality models. *Journal of Environmental Engineering* **116**, 250-268.

Presents several statistical methods for model evaluation with applications illustrating their use. Discusses the role of a goodness-of-fit test. Encourages the use of graphical techniques.

Keywords: water research, statistics, prediction, prediction measures, general, goodness-of--fit, testing, graphics

247. Reynolds, M.R., Burkhart, H.E. & Daniels, R.F. (1981). Procedures for statistical validation of stochastic simulation models. *Forest Science* **27**, 349-364.

Presents parametric and nonparametric methods for testing that the observed and simulated response distributions are the same.

Keywords: statistics, ecology, testing, forestry, output, goodness-of-fit

248. Rice, J.A. (1988). *Mathematical Statistics and Data Analysis*, Pacific Grove, CA:Wadsworth and Brooks.

Briefly describes chi-squared goodness-of-fit tests and shows some graphics (e.g., hanging rootograms and hanging chigrams) useful for evaluating fit.

Keywords: goodness-of-fit, graphics, testing

249. Richels, R. (1981). Building good models is not enough. *Interfaces* **11**, 48-54.

Asserts that there is a growing realization that independent or third-party assessment is critical if policy makers are to have confidence in model results.

Keywords: management, energy, documentation, verification, third party

250. Rosenbaum, P.R. & Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* **39**, 33-38.

Proposes an algorithm for pairing observations; it is based on propensity scores and Mahalanobis distances. Could be used to create predicted/observed matched pairs.

Keywords: statistics, design, matching, biomedical, logistic regression, prediction measures

251. Rouhani, S. & Wackernagel, H. (1990). Multivariate geostatistical approach to space-time data analysis. *Water Resources Research* **26**, 585-591.

For describing the structure of time-rich, space-poor hydrology data, proposes a multivariate time series analysis. Analysis based on a cross-variogram between two locations.

Keywords: water research, spatial model, cross-correlation, empirical, hydrology, nonstationary, time series, structure, cluster analysis, cross-correlation

252. Rubin, D.B. (1981). The Bayesian bootstrap. *Annals of Statistics* **9**, 130-134.

Creates a Bayesian analog of the bootstrap procedure, describes similarities to the usual frequentist bootstrap procedure, and cautions that both procedures are sensitive to model assumptions.

Keywords: statistics, bootstrap, uncertainty, sensitivity, Bayesian

253. Ruff, R.E., Nitz, K.C., Ludwig, F.L., Bhumralkar, C.M., Shannon, J.D., Sheih, C.M., Lee, I.Y., Kumar, R. & McNaughton, D.J. (1985). Evaluation of three regional air quality models. *Atmospheric Environment* **19**, 1103-1115.

Three air quality models were evaluated for face (scientific) validity and for ability to predict real data. Because of autocorrelation, the data were preprocessed to create essentially independent observations. Potential spatial correlation was not mentioned.

Keywords: structure, output, input, sensitivity, substructure, prediction, prediction measures, air pollution, mechanistic, face validity, serial correlation

254. Russell, A.G., McCue, K.F. & Cass, G.R. (1988). Mathematical modeling of the formation of nitrogen-containing air pollutants. 1. Evaluation of an Eulerian photochemical model. *Environmental Science and Technology* **22**, 263-271.
Model is numerical solution of diffusion equations. It predicts time-space concentrations of pollutants. Initial conditions and boundary conditions are based on the literature. Various numerical and graphical comparisons of predicted to observed. Apparently ignores potential spatial-temporal correlations.
Keywords: structure, output, prediction, mechanistic, numerical model, calibration, air pollution, graphics, prediction measures

255. Russell, A.G., McCue, K.F. & Cass, G.R. (1988). Mathematical modeling of the formation of nitrogen-containing pollutants. 2. Evaluation of the effect of emission controls. *Environmental Science and Technology* **22**, 1336-1347.
Example of how a model is used for setting control levels or for policy decision. Presents tabular and graphical comparisons of pollution levels for various assumed control measures.
Keywords: air pollution, control, mechanistic, prediction, graphics

256. Sacks, J., Welch, W.J., Michell, T.J. & Wynn, H.P. (1989). Design and analysis of computer experiments (with discussion). *Statistical Science* **4**, 409-435.
Suggests a cubic spline interpolating function for approximating a complicated deterministic computer spatial model. The model can be used to explore the response surface, to identify especially influential input variables, etc. (but not for extrapolation), and as a numerical analysis method for computing complicated integrals in Bayesian analyses. With discussion.
Keywords: computer model, statistics, design, spatial model, sensitivity, mechanistic, empirical, deterministic, computer experiment

257. Schervish, M.J. (1989). A general method for comparing probability assessors. *Annals of Statistics* **17**, 1856-1879.
Presents a general method for creating scoring rules which can be used to compare probability assessors.
Keywords: statistics, Bayesian, forecasting, output, prediction measures, expert opinion

258. Schmidt, B. (1987). What does simulation do? Simulation's place in the scientific method. *Systems Analysis, Modelling, Simulation* **4**, 193-211.
Summary of the modelling process with a glossary of terms. Includes examples of differential equations models.
Keywords: structure, simulation, mechanistic, computer model, general

259. Seo, D. & Krajewski, W.F. (1990). Stochastic interpolation of rainfall data from rain gages and radar using cokriging 2: Results. *Water Resources Research* **26**, 915-924.
Second of a two-part series. Methods of first paper are evaluated. Many graphical and numerical analyses shown.
Keywords: water research, hydrology, prediction, prediction measures, input, uncertainty, graphics

260. Shaffer, A.R. & Elsberry, R.L. (1982). A statistical-climatological tropical cyclone track prediction technique using an EOF representation of the synoptic forcing. *Monthly Weather Review* **110**, 1945-1954.
Studies a new method, based on principle components stepwise regression, for tracking tropical cyclone movement. Compares forecasting errors to the standard method on an independent data set for forecasts made 1, 2, and 3 days ahead.
Keywords: meteorology, structure, output, prediction, empirical, hold-out data, stepwise regression

261. Shapiro, L.J. (1984). Sampling errors in statistical models of tropical cyclone motion: a comparison of predictor screening and EOF techniques. *Monthly Weather Review* **112**, 1378-1388.

Uses principle components to reduce the number of grid- points used as predictors. Compares models developed from preselected and ordered predictors and from a forward stepwise variable selection technique in a manner that takes account of spatial and serial correlations. Determines the number of effectively independent observations.

Keywords: meteorology, regression, empirical, serial correlation, statistics, stepwise regression, spatial correlation

262. Shapiro, S.L. & Teukolsky, S.A. (1988). Building black holes: supercomputer cinema. *Science* **241**, 421-425.

Description of an intriguing computer model for simulating the collapse of an unstable star cluster to a supermassive black hole at the center of a galaxy.

Keywords: computer model, structure, discovery, mechanistic, physics, explanation

263. Shibata, R. (1986). Regression variables, selection of. In *Encyclopedia of statistical sciences - Volume 7*, Eds. Kotz, S., Johnson, N.L. & Read, C.B., pp. 709-714. New York, New York: John Wiley & Sons.

Reviews popular criterion procedures for chosing the most parsimonious multiple regression model that fits the data reasonably well. Criteria mentioned include: multiple correlation coefficient, forward and backward stepwise regression, final prediction error, Akaike's information criterior, Mallow's Cp criterion, the Hannan and Quinn criterion, and the Bayes Information Criterion.

Keywords: statistics, regression, selection, input, structure, review

264. Sims, C.A. (1988). Uncertainty across models. *American Economic Review* **78**, 163-167.

Discusses three strategies, (i) locating dimensions of uncertainty on which the data are uninformative, (ii) testing the models' fit to data, and (iii) weighting schemes for combining models.

Keywords: economics, Bayesian, specification, forecasting, testing, uncertainty

265. Singh, S.V. & Kripalani, R.H. (1986). Potential predictability of lower-tropospheric monsoon circulation and rainfall over India. *Monthly Weather Review* **114**, 758-763.

Calculates the ratio of variation between years (signal) to the variation within years (noise). Tests for significance using the F statistic calculated using the estimated time between effectively independent observations.

Keywords: meteorology, serial correlation, prediction, prediction measures, inherent error

266. Singh, V.P. & Yu, F.X. (1987). A mathematical model for border irrigation III. Evaluation of models. *Irrigation Science* **8**, 191-213.

Models are evaluated against 10 data sets. Looked at response versus time, but there was no mention of potential serial correlation.

Keywords: agriculture, prediction measures, output, prediction

267. Smith, G. & Galligan, D.T. (1988). Mathematical models of the population biology of ostertagia ostertagi and teladorsagia circumcincta, and the economic evaluation of disease control strategies. *Veterinary Parasitology* **27**, 73-83.

Uses a deterministic population model to simulate the results of alternative disease control strategies. Calibration by nonlinear least squares. Evaluation by plotting predicted and observed against time.

Keywords: deterministic, biomedical, graphics, explanation, control

268. Smith, J.Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting* **4**, 283-291.
Suggests diagnostics for checking the predictive distribution. Although a formal goodness--of-fit test could be conducted using the statistics suggested, the authors focus on graphical diagnostics.
Keywords: forecasting, prediction, prediction measures, output, time series, serial correlation, diagnostics, dynamic, structure, graphics

269. Snapinn, S.M. & Knoke, J.D. (1989). Estimation of error rates in discriminant analysis with selection of variables. *Biometrics* **45**, 289-299.
Describes an alternative to the bootstrap procedure for estimating error rates. The method requires normality but requires little computational effort.
Keywords: classification, statistics, bootstrap, selection, structure, output, empirical

270. Snee, R.D. (1990). Statistical thinking and its contribution to total quality. *American Statistician* **44**, 116-121.
Thoughtful overview of efforts to achieve total quality in any process. The terminology, outline, and strategies fit well the model building/model validation processes.
Keywords: quality control, statistics, review

271. Solow, A.R. (1989). Is it getting stuffy in here, or is it just my imagination?. *Chance:New Directions for Statistics and Computing* **2**, 40-46.
Reviews the status of climate modelling as it relates to the greenhouse effect and global warming.
Keywords: earth science, uncertainty, prediction, mechanistic, feedback model, discovery

272. Son, M.S., Holbert, D. & Hamdy, H. (1987). Bootstrap forecasts versus conventional forecasts: a comparison for a second order autoregressive model. *American Statistical Association Proceedings of the Business and Economic Statistics Section* 723-727.
Compares the bootstrap and conventional methods for determining forecast reliability.
Keywords: bootstrap, time series, statistics, uncertainty, forecasting

273. Spiegelman, C.H. (1986). Two pitfalls of using standard regression diagnostics when both X and Y have measurement error. *American Statistician* **40**, 245-248.
In simple regression when both Y and X have measurement error, the author's examples show that systematic model departures and heteroscedasticity may not be detectable with standard regression diagnostics.
Keywords: structure, statistics, regression

274. Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics* **29**, 143-151.
Describes the Latin hypercube sampling procedure and a method for estimating the standard error of the simulation estimate (correction: Technometrics 1990 32:367)
Keywords: statistics, engineering, design, sensitivity, simulation, computer experiment

275. Stocker, M. & Noakes, D.J. (1988). Evaluating forecasting procedures for predicting Pacific herring (Clupea harengus pallasi) recruitment in British Columbia. *Canadian Journal of Fisheries and Aquatic Science* **45**, 928-935.
Four models for forecasting Pacific herring population characteristics are compared against real data.
Keywords: ecology, time series, empirical, mechanistic, prediction, prediction measures, output, forecasting, selection

276. Stockton, D.J. & Struckmeyer, C.S. (1989). Tests of the specification and predictive accuracy of nonnested models of inflation. *The Review of Economics and Statistics* **71**, 275-283.

Three models for explaining aggregate price inflation are compared to data and evaluated for model specification and predictive accuracy.

Keywords: economics, specification, prediction, prediction measures, structure

277. Stone, A. & Seip, H.M. (1989). Mathematical models and their role in understanding water acidification: An evaluation using Birkenes model as an example. *Ambio* **18**, 192-199.

Assesses the structure, substructures, stability of calibration, sensitivity, and compares observed to predicted.

Keywords: hydrology, tracer, mechanistic, calibration, sensitivity, prediction, output, substructure, structure

278. Stoto, M.A. (1988). Dealing with uncertainty: statistics for an aging population. *American Statistician* **42**, 103-110.

Reviews the kinds of uncertainties that appear in statistical data used to guide policy. Recommends confidence or credible intervals. Recommends against using words such as likely, probably, and rarely because these words mean different things to different people.

Keywords: output, prediction, uncertainty, biomedical, demography, expert opinion

279. Stunder, M. & Sethuraman, S. (1986). A statistical evaluation and comparison of coastal point source dispersion models. *Atmospheric Environment* **20**, 301-315.

Two complicated dispersion models are evaluated against measurements of SO_2 emitted from a point source. Mobile, in-situ, and fixed monitors were used. Evaluation uses a variety of statistical measures and scatter plots.

Keywords: structure, output, prediction, mechanistic, empirical, substructure, sensitivity measures, prediction measures, air pollution

280. Swartzman, G.L. & Kaluzny, S.P. (1987). *Ecological Simulation Primer*, New York:Macmillan Publishing Company.

Discusses linear differential equation simulation models, computer coding methods, nonlinear models, steps in model development, suggestions for model documentation and coding, stochastic models, model evaluations and analyses, spatial aspects, and some case studies. Includes a glossary of 10 pages.

Keywords: ecology, simulation, design, sensitivity, documentation, prediction measures, mechanistic, input, structure, output, general, deterministic, stochastic

281. Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science* 240, 1285-1293.

Proposes a measure of accuracy for assessing a diagnostic test in medicine. The measure of accuracy is neither dependent upon the prevalence of the disease in question nor on the decision criteria to be used.

Keywords: prediction measures, sensitivity measures, permutation test, power, structure, output, testing, statistics, biomedical

282. Switzer, P. (1989). Model evaluation. *Sixth Annual EPA Conference on Statistics* pp. 7-8.

Thought provoking list of important issues in model evaluation.

Keywords: general

283. Tawn, J.A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika* **75**, 397-415.
Mathematical analysis of environmental extreme value data where there is a need for models of dependence between extremes at different spatial locations; e.g., at two neighboring seaports. Discusses tests for independence between bivariate extreme values and for discriminating between probability models.
Keywords: statistics, extreme events, meteorology, structure, selection, output, testing

284. Terasvirta, T. & Mellin, I. (1986). Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics* **13**, 159-171.
Proposes critical values of a long list of model selection criteria. Focuses on nested models when perhaps none of the models are correct. Uses squared error loss.
Keywords: input, output, empirical, selection, uncertainty, regression, serial correlation, testing

285. Tesche, T.W., Haney, J.L. & Morris, R.E. (1987). Performance evaluation of four grid-based dispersion models in complex terrain. *Atmospheric Environment* **21**, 233-256.
Evaluates a dispersion model by pairing the observations in time and space and calculating various measures. Discusses transfer matrix methods by which the contribution to the concentration at a receptor is attributed to each of the potential sources.
Keywords: prediction, prediction measures, mechanistic, tracer, air pollution, sensitivity, causal effects

286. Thiebaux, H.J. & Zwiers, F.W. (1984). The interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology* **23**, 800-811.
Discusses the determination of the effective sample size when the data are temporally or spatially correlated. Criticizes the concepts of "equivalent number of degrees of freedom" and "number of effective degrees of freedom." Discusses t-test approaches when correlations exist.
Keywords: statistics, prediction measures, serial correlation, empirical, time series, testing, spatial correlation

287. Thomann, R.V. (1982). Verification of water quality models. *Journal of Environmental Engineering* **108**, 923-940.
Lists the main components of mathematical modelling and the steps to follow in model validation, including a continuing model post-audit (surveillance). Emphasizes the value of statistical assessments.
Keywords: structure, output, prediction measures, prediction, mechanistic, water research, calibration, surveillance, graphics

288. Thompson, J.R. (1989). *Empirical Model Building*, New York:John Wiley and Sons.
Includes differential equations models, simulation-based techniques, exploratory & graphical data analysis, quality control, and computer intensive methods. Not devoted to the usual least squares regression methods often used in empirical modelling.
Keywords: mathematics, general, structure, numerical analysis, computer model, empirical, quality control, simulation, stochastic, graphics

289. Thrall, A.D. (1989). Model evaluation: A neglected issue?. In *Challenges for the '90s*, Eds. Bailar, B.A. & Leone, F.C., pp. 42-45. Alexandria, VA: American Statistical Association.
In a brochure produced for the 150th anniversary of the American Statistical Association.

Identifies model evaluation as one of the major challenges facing statisticians. Describes the potential statistical contributions to model validation.
Keywords: ecology, biomedical, statistics

290. Trenberth, K.E. (1984). Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations. *Monthly Weather Review* **112**, 2359-2368.
Derives and evaluates a formula for determining the time between effectively independent observations for serially correlated data. Describes how the formula can be used in moment estimation and in testing.
Keywords: meteorology, serial correlation, statistics, testing

291. Tsang, H.Y. & Bacon, D.W. (1980). Detection of unsuspected feedback in linear dynamic systems. *Technometrics* **22**, 509-516.
Highly technical paper proposing a diagnostic procedure for detecting a feedback mechanism. Applied toward modelling the dynamic behavior of an industrial petroleum fractionation process.
Keywords: engineering, dynamic, structure, diagnostics, cross-correlation, time series

292. Unkelbach, H. & Passing, H. (1990). Quality assurance of statistical software. *Statistical Software Newsletter* **15**, 49-55.
Reviews techniques for assuring the quality of computer programs. Provides a list of important documents required to make the model usable.
Keywords: statistics, software, access, quality control

293. van der Laan, P. & Verdooren, L.R. (1989). Selection of populations: an overview and some recent results. *Biometrical Journal* **31**, 383-420.
Review paper containing 130+ references. Describes both the indifference zone and subset selection approaches.
Keywords: ranking and selection, statistics, selection, review, historical, bibliography

294. Vandenbroucke, J.P. (1987). Should we abandon statistical modeling altogether. *American Journal of Epidemiology* **126**, 10-13.
Suggests scientific conclusions be based on simple stratified analyses whenever possible and that models should be checked against the simpler analyses.
Keywords: biomedical, empirical, structure

295. Venkatram, A. (1982). A framework for evaluating air quality models. *Boundary-Layer Meteorology* **24**, 371-385.
Models the association between the predicted outcome and observed output and suggests associated visual and numerical measures of misspecification.
Keywords: air pollution, mechanistic, prediction, missing data, structure, output, specification, misspecification

296. Venkatram, A. & Karamchandani, P.K. (1988). Testing a comprehensive acid deposition model. *Atmospheric Environment* **22**, 737-747.
Predictions of amount of chemical in acid rain are compared to observations. The scientific plausibility of an important submodel is studied.
Keywords: air pollution, prediction, prediction measures, output, uncertainty, mechanistic, substructure

297. Versar, Inc. (1989). Guidance in assessing the validation of exposure models, Draft Final Report, EPA Contract No. 68-02-4254-Task 206, Sept. 30, 1989, Versar, Inc., 6850 Versar Center, Springfield, VA 22151. (UnPub)
This is an overview of model validation, especially as applied to the determination of levels of pollutants in surface water, ground water, and air. A glossary is provided. Describes major steps in the validation process, including an interesting section on post-development validation.
Keywords: general, environmental, input, structure, output, calibration, sensitivity, statistics, uncertainty, review, surveillance

298. Vesely, W.E. & Masmuson, D.M. (1984). Uncertainties in nuclear probabilistic risk analyses. *Risk Analysis* **4**, 313-322.
Identifies the uncertainties associated with a probabilistic risk analysis (PRA) for nuclear power plant accidents and the primary needs in PRA uncertainty analyses. List various utilizations of PRAs.
Keywords: uncertainty, energy, sensitivity, structure, input, calibration

299. Vincent, J., Waters, A. & Sinclair, J. (1988). *Software Quality Assurance-Volume II: A Program Guide*, Englewood Cliffs, NJ:Prentice Hall.
Especially useful for a list of criteria, with definitions.
Keywords: quality control, verification, software

300. Warnes, J.J. & Ripley, B.D. (1987). Problems with likelihood estimation of covariance function of spatial Gaussian processes. *Biometrika* **74**, 640-642.
Warns of difficulties to be encountered when trying to estimate the covariance parameters of a spatial Gaussian process using the maximum likelihood technique.
Keywords: statistics, spatial correlation, time series, spatial model, calibration

301. Wegman, E.J. (1988). On randomness, determinism, and computability. *Journal of statistical planning and inference* **20**, 279-294.
Views on the definition of randomness and the relationship of chaotic dynamical systems to randomness.
Keywords: statistics, mathematics, stochastic, mechanistic, deterministic

302. Weil, J.C. (1985). Updating applied diffusion models. *Journal of Climate and Applied Meteorology* **24**, 1111-1130.
This paper is an overview of the recommendations arising from an AMS/EPA Workshop held in Florida in January, 1984. Discusses substructures of the overall model, uncertainty, and inherent variability.
Keywords: air pollution, inherent error, prediction, prediction measures, output, mechanistic, substructure, graphics, uncertainty, review

303. Weisberg, S. (1985). *Applied Linear Regression*, New York:John Wiley & Sons, Inc..
Textbook on regression analysis. It is a particularly good reference for variable selection methods and diagnostic techniques.
Keywords: statistics, regression, empirical, diagnostics, graphics, selection, prediction, residual analysis

304. White, W.H., Seigneur, C., Heinold, D.W., Eltgroth, M.W., Richards, L.W., Roberts, P.T., Bhardwaja, P.S., Conner, W.D. & Wilson, W.E., Jr. (1985). Predicting the visibility of chimney plumes: an intercomparison of four models with observations at a well-controlled power plant. *Atmospheric Environment* **19**, 515-528.

In this assessment of some plume visibility models, the authors investigate the uncertainties in the observations, uncertainties in components of input and of substructures, the extent of observation-prediction differences, and the importance of breaking the overall model into testable substructures, especially when the data are limited in number.
Keywords: air pollution, input, structure, output, substructure, graphics, data quality, sensitivity, mechanistic, uncertainty, inherent error

305. Whitner, R.B. & Balci, O. (1989). Guidelines for selecting and using simulation model verification techniques. Technical Report TR-89-17, Department of Computer Science, Virginia Polytechnic Institute and State University, Blackburg,VA 24061. (UnPub)

A systematic review of techniques for computer program verification, testing, debugging, and quality assurance. The techniques are classified as informal, static, dynamic, symbolic, constraint, or formal.
Keywords: simulation, structure, face validity, substructure, software

306. Wigley, T.M.L. & Santer, B.D. (1990). Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *Journal of Geophysical Research* **95**, 851-865.

Suggests a variety of test statistics for comparing means, variances, and patterns of spatial fields in situations where the data are possibly affected by temporal and spatial autocorrelations.
Keywords: earth science, spatial correlation, time series, output, prediction, prediction measures, testing, goodness-of-fit, serial correlation

307. Wilks, D. (1987). Sampling distribution of the time between effectively independent samples. *Monthly Weather Review* **115**, 740-743.

Using Monte Carlo simulation, investigates the mean, standard error, and sampling distribution of Trenberth's (1984) formula for estimating the time between effectively independent samples for an autoregressive process. The simulations used autoregressive order 1, 2, or 3.
Keywords: statistics, time series, serial correlation

308. Will, C.M. (1990). General relativity at 75: how right was Einstein?. *Science* **250**, 770-776.

Reviews efforts to compare and contrast observable predictions of Einstein's general relativity with the predictions of alternative theories of gravity. Because the predictions of general relativity are fixed, containing no adjustable constants, a verified discrepancy between observation and prediction would discredit the theory. So far, general relativity has survived every test.
Keywords: physics, mechanistic, testing, prediction measures

309. Willmott, C.J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* **63**, 1309-1313.

Argues that the correlation coefficient for paired observed and predicted values should not be used as a model performance measure but that scatter plots are important. Proposes some measures for performance assessment.
Keywords: prediction, prediction measures, output, statistics, meteorology, graphics

310. Wold, H. (1988). Specification, Predictor. In *Encyclopedia of statistical sciences - Volume 8*, Eds. Kotz, S., Johnson, N.L. & Read, C.B., pp. 587-599. New York: John Wiley & Sons.

Discusses prediction models in econometrics and the behavioral sciences; e.g. path models. Contrasts maximum likelihood to least squares.
Keywords: behavioral science, economics, selection, input, prediction, calibration, jackknife, uncertainty, testing

311. Wooley, R.N. & Pidd, M. (1981). Problem structuring - A literature review. *Journal of the Operational Research Society* **32**, 197-206.
Problem structuring is the process by which the initially presented set of conditions is translated into a set of problems, issues and questions sufficiently well-defined to allow specific research action. Proposes four streams of thought on problem structuring, the checklist, the definitions, science research, and people. 58 references.
Keywords: engineering, management, causal effects, general, bibliography

312. Wu, C.FJ. (1986). Jackknife, bootstrap and other resampling methods in regression analysis with comments and reply. *Annals of Statistics* **14**, 1261-1350.
Suggests an extension of the jackknife method for estimating standard errors and confidence intervals in linear models and generalized linear models problems. Points out potential difficulties with the bootstrap procedure. With discussion.
Keywords: structure, statistics, regression, nonlinear regression, bootstrap, jackknife, uncertainty

313. Zeigler, B.P. (1976). *Theory of Modelling and Simulation*, New York:John Wiley & Sons, Inc..
Systematic discussion of modeling, uses for models, and assessment of models. A broad spectrum of topics is covered, from model conception to model evaluation against real data.
Keywords: general, computer model, simulation, input, output, structure

314. Zimmermann, H.J. (1980). Testability and meaning of mathematical models in social sciences. *Mathematical Modelling* **1**, 123-139.
Reviews the terminology and strategies of mathematical model building. Gives various definitions of models and types of models. Lists possible uses of models. Describes the critical rationalism of Karl Popper and other philosophical aspects of modeling. Discusses three different sorts of testing: 1. Testing while the model is being built - a priori testing. 2. Testing on the basis of the outputs of the analysis. 3. Testing by implementation -surveillance.
Keywords: general, philosophy, mathematics, testing, behavioral science

315. Zwiers, F.W. (1987). A potential predictability study conducted with an atmospheric general circulation model. *Monthly Weather Review* **115**, 2957-2974.
Studies potential predictability and inherent variability (or natural variability), in this case climate noise, using signal/noise concepts.
Keywords: meteorology, serial correlation, prediction, output, time series, mechanistic, inherent error, testing

316. Zwiers, F.W. & Thiebaux, H.J. (1987). Statistical considerations for climate experiments. Part I: Scalar tests. *Journal of Climate and Applied Meteorology* **26**, 464-476.
Examines several scalar tests for differences of means and variances between predicted and observed. Applies tests which take the serial correlation structure of the time series into account. Favors use of the time between effectively independent observations over using a standard test adjusted by an equivalent sample size calculation.
Keywords: meteorology, time series, serial correlation, prediction, prediction measures, statistics, testing

APPENDIX A. GLOSSARY

# indicates a key word

access# - pertains to a computer model and the ease with which it can be obtained, understood, and used

aggregation - treating as a whole a collection of units (cases or variables) that are somewhat loosely associated with one another; often the association is in time or space

assessment - model validation conducted by a third party

calibration# - an adjustment (often determined by statistical estimation techniques such as least squares regression) to choose parameter values so that the output estimates are close to the measured system output values. The calibrated model is technically valid only for a particular site and data base. For a dynamic model, calibration has been defined as the process of changing model parameters to obtain an improved fit of model output variables to data time traces (Swartzman and Kaluzny 1987). Also called model tuning. See identifiable model.

computer experiment# - where the computer model is run at input variable and parameter values chosen by the experimenter with an objective such as (i) predict the response at observed inputs, (ii) optimize a function of the output variables, or (iii) tune the computer code to physical data. This key word was also given to articles using simple, surrogate computer models for experimentation instead of a large computer model that is expensive to run.

computer model# - a numerical model converted to computer code so that the input and parameter values are entered into a computer and the program determines the associated output values.

control# - as a use for the model - to suggest target values for input variables in order to optimize some objective function of output variables

descriptive model - the model simply reproduces input-output responses for the available experimental data; the fit of the model is not directed at any specific purpose for the model.

design# - indicates an article pertaining to the design of an experiment or study conducted for purposes of model evaluation, calibration, computer program verification, etc.
- as a purpose for the model - the model is to be used to design a subsequent experiment or to design a product, building, etc.

empirical# - as in empirical model, a model determined by statistically fitting equations to data; in contrast to a mechanistic model and a descriptive model.

face validity# - the surface or initial impression of a model's realism obtained by asking people who know the real system; depends on expert opinion

identifiable model - a model for which all unknown parameters can be deduced from plausibility and internal consistency arguments or can be estimated by formal statistical calibration techniques such as least-squares or maximum likelihood. The calibration produces a model that is uniquely identifiable, non-uniquely identifiable, or unidentifiable.

inherent error# - Consider a conceptual number of repeats of the same "experiment," holding the environment (parameters) and external conditions (boundary and initial conditions) fixed. Then the difference between individual observed outputs and the mean output indicates the inherent error. This is philosophically distinct from an error that represents in some sense inadequacy in either a measuring procedure or in the model. For example, in air quality assessment, inherent error arises from the fact that concentration values, measured in individual realizations of nominally the same experiment will be different because of the stochastic nature of turbulent dispersion. Inherent error would remain even if both error-free model and input data could be developed. Its magnitude represents the best a model can do.

initial conditions - the spatial profile $s(t_o)$ of the output at time t_o, information required in order to use a dynamic model to calculate future or past behavior

internal validity - the model (i) contains no mathematical or logical contradictions, and (ii) uses valid algorithms that are appropriate, accurate, stable, convergent, and produce acceptably small round-off errors.

input# - as in the choice, measurement, or quality of input data.

lumping - aggregating a continuous (in time or space) model into a discrete model

matching# - as in matching observed data with model-predicted values

mechanistic# - as in mechanistic model, a representation of the physical or mechanistic theory governing the system; in contrast to an empirical model.

numerical validity - assessment of the numerical aspects of converting a complicated mathematical model into an approximate model or a model capable of being programmed on a computer

output# - as in

model output variables - variables determined by the model; they are dependent on the input variables, parameters, and model structure, or

system output variables - variables of interest that can be measured directly or indirectly; system output variable values are thought to be related to or caused by specified input variables.

parameters -

- of the mathematical model - numerical constants that describe the environment within which the model is operating. The parameters index the collection of possible models. The act of choosing values for the parameters amounts to choosing a model from the collection. In using a computer model, the parameters will be fixed and the model run with different values for the input variables to determine or estimate

the associated values of the output variables. In some of the modeling literature, the distinction between parameters and input variables is fuzzy, but generally the parameters are true constants that cannot be manipulated in reality whereas the input variables may be manipulated.
- of the probability distribution of data - quantities that determine the specific distribution within a well-defined class of possible distributions

prediction measures# - indicates an article that uses or discusses measures suitable for evaluation; that is, suitable for comparing model predicted output values to observed values

performance analysis# - similar to model assessment, but performance analysis has the added element of uncertainty due to scenario (e.g., earthquake near a nuclear power plant), the range and distribution of appropriate input variables and parameters that can alter the model

selection# - as in
- variable selection - choosing only effective input variables (predictors) for (i) reducing the prediction error or (ii) reducing the number of input variables required to achieve a stable model; or
- model selection - selecting the best from among several competing models, where "best" depends upon the purpose for the model

sensitivity# - refers to
global sensitivity - variation of model behavior due to possibly large perturbations of the parameters; or
local sensitivity - variation of model behavior due to small perturbations of the parameters; often determined by taking derivatives of the mathematical model with respect to the parameters; or
sensitivity analysis - determination of the extent to which model behavior varies as model parameters are varied

side conditions - initial conditions or boundary conditions or any other restriction on the possible inputs, parameters, structure or outputs

software# - pertains to software quality and software engineering, the latter being the study of the various aspects and phases of the life cycle of a piece of computer software

specification# - the choice of model structure

structure# - generic word for the relationship between the output variables and the input variables/parameters. For a mathematical model, the structure is the set of equations relating the output to input variables/parameters.

surrogate# - as in
- surrogate model - an approximate model used in some instances (e.g., computer experiments) instead of the exact model; or
- surrogate variable - a variable used in place of (i) the specific input variable required by the model or (ii) the specific output variable predicted by the model

surveillance# - either

- as a use for the model - to help determine the best plan for watching the system over time/space to detect trends, variations, or unusual occurrences; or
- the protocol to follow in the continuing evaluation of a model as it is applied to a variety of goals in a variety of environments

third party# - neither the model builder nor the model user

APPENDIX B. KEY WORD INDEX

ACKNOWLEDGEMENT: This work was supported in part by U.S. Environmental Protection Agency Cooperative Agreement R-815142, D. Goodman, Project Manager. Andrea Easter-Pilcher provided able assistance in locating relevant articles. The author benefitted from discussions with Drs. Robert Boik, Warren Esty, and Daniel Goodman.

Received February 1991; March 1991.

Recommended Anonymously.