International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Performance Evaluation of Environmental Models

Neil D. Bennett[1], Barry F.W. Croke[1,2], Anthony J. Jakeman[1], Lachlan T.H. Newham[1] and John P. Norton[1]

[1] *Integrated Catchment Assessment and Management Centre, Fenner School of Environment and Society, The Australian National University, Canberra ACT 2600, Australia*

[2] *Department of Mathematics, The Australian National University, Canberra ACT 2600, Australia (barry.croke@anu.edu.au)*

Abstract: For environmental models to be effectively used for management and decision making purposes it is necessary to have confidence in their performance. This paper reviews techniques available for performance evaluation of environmental models. Both quantitative and qualitative performance evaluation are considered and application recommendations for environmental models are provided.

Keywords: Model development; model testing; performance indicators; sensitivity analysis

1 INTRODUCTION

Environmental models are increasingly and extensively used in research, management and decision making. Hence, the importance of assessing confidence in the outputs of such models has increased. The question of model evaluation or how well a model represents the system under study has resulted in many different approaches and much debate on the appropriateness of these techniques. The best method will depend on the type of model, the data and aims of modelling, and multiple methods may be needed for the best understanding and decision support. This paper reviews evaluation methods and criteria available for environmental models. Although the primary applications under consideration are environmental, methods that have been developed in other fields are also included.

Modelling is an essential component of environmental management and science, necessary for understanding and representing environmental systems as well as increasing confidence in management decisions. Currently modelling is used across many environmental fields – hydrology, oceanography, climate change to name a few. In each of these fields many different types of model can be used, which will affect how the performance of a model is evaluated. Jakeman et al (2006) separate models into the following model families:

- Empirical: data-based, statistical models where a versatile structure is assumed with minimal assumptions. Examples include cluster analysis, time series models and regression analysis.
- Stochastic: general form models that have a standard structure allowing the incorporation of previous knowledge and uncertainty e.g. state space and hidden Markov models.
- Specific process/theory based models (usually called deterministic: have a set structure specific to the process and justified by prior theory.
- Conceptual models: create a structure based on assumed cause-effect links e.g. Bayesian decision networks and compartmental models.

- Agent-based models: locally structured models that allow for emergent, unpredicted behaviour.
- Rule-based models: a group of models that use rules to represent and simulate the interactions/behaviour between discrete events and decisions. The model can then map the probability of different outcomes including decision trees and expert systems.
- Models incorporating dynamics: a spectrum of models that can give time-spread responses to an input at any instant. The spectrum includes discrete event/state, lumped dynamical, distributed and delay-differential infinite-state-dimensional models.
- Spatial models: including region based, polygon based and pseudo-continuous spatial models.

Both the intended application of the model and its main features will affect the overall behaviour of the model. Therefore, when selecting a performance evaluation procedure, it is necessary to take these details into account.

Verification and validation of models are core components of the modelling process and significant research has focussed on these topics, including philosophical debates to differentiate verification from validation (Jakeman et al 2006, Oreskes et al, 1994). This paper is not concerned with such debate; instead it deals with the performance evaluation methods assessing environmental models, whether these methods and criteria are used for verification, validation or calibration. At its highest level, performance evaluation can be split into two categories: quantitative methods that test the model against measured data; and qualitative testing, which does not require independent data to evaluate the model.

2 QUANTITATIVE TESTING

Many different methods have been introduced to evaluate the performance of a model. In hydrology previous work has been completed on general modelling frameworks that consider performance criteria as part of the framework (Jakeman et al. 2006 and Wagener et al. 2001), while studies have also been completed that focus explicitly on performance criteria. Moriasi et al. (2007) produced guidelines for systematic model evaluation, including a list of recommended evaluation techniques and performance metrics. Dawson et al. (2007) has also produced a comprehensive list of metrics that can be applied to hydrological forecasting models as well as a web-based toolbox, HydroTest, that can be used to calculate the metrics.

Different techniques that can be used to test models are:

- Data division methods
  - Cross validation (e.g. Kohavi, 1995, Klemes, 1986)
  - Bootstrapping
- Direct comparison methods
  - e.g. plots of data points and frequency distributions
  - statistical metrics comparing modelled with observed values
  - regression of observed and modelled values, or sum and difference of observed and modelled values (Kleijen et al., 1998)
- Residual methods
  - Graphical methods: e.g. residual and Q-Q plots
  - Numerical methods: e.g. bias, mean square error, mean absolute error, maximum absolute error, error in peak, relative volume error
  - Use of transformations to handle heteroscedasticity in the residuals: e.g. Box-Cox transformation (note care is needed as deficiencies in the model structure may contribute to the heteroscedasticity in the residuals)
  - Impact of variations in uncertainty through the data: e.g. Heteroscedastic Maximum Likelihood Estimator to allow for variation in the divergence of the residuals (Sorooshian and Dracup, 1980). See also Croke (2007 and 2009) for a more general form, including allowing for serial correlation.
  - Application of methods listed here on subsets of the data – selection of subsets based on different aspects of the system response (e.g. Boyle et al. 2000) or using a moving window (e.g. Choi and Beven, 2005) for example.

  - Information criteria: Akaike Information Criterion (Akaike, 1974); Bayesian Information Criterion (Schwartz, 1978); Young Information Criterion (Young, 2001, Taylor et al., 2007)
  - Model efficiency: (e.g. coefficient of determination, correlation coefficient)
  - Parameter error and identifiability: (e.g. Jakeman and Hornberger, 1993; Young et al., 1980; Checchi et al., 2006, Wagener et al., 2003, Härdle et al., 2003)
- Transformation methods
  - Fourier and Wavelet transforms to convert residuals to the frequency domain (e.g. Lane, 2007, Chou, 2007)
- Spatial methods
  - Global and local spatial methods: global methods act over the entire spatial domain (ignore any spatial characteristics) while local methods are applied over restricted domains.
  - Grouping spatial methods: e.g. defining homogeneous regions (Wealands, 2005); multi-scale approaches, empirical orthogonal functions (Hannachi et al., 2007).
  - Categorical spatial methods: e.g. confusion matrix (Congalton, 1991); kappa statistic; fuzzy maps (Wealands, 2005).
- Multi-criteria methods: typically involve the concept of a Pareto optimal set (Gupta et al., 1997; Yapo et al., 1998)
- Diagnostic based evaluation methods: consider the information contained in the data (Gupta et al., 2008), exploring impact of model structure (Clark et al. 2008).

3 QUALITATIVE TESTING

In the case when data is unavailable for quantitative testing qualitative testing, which aims to provide a consistent means to compare model performance, is the only form of testing possible. It is, however, highly beneficial for performance testing even when data is available for quantitative testing and should be included within a standard model development routine (Jakeman et al, 2006).

The core component of qualitative testing is a face validation or Turing test, which calls for the analysis of the output and operation of the model to see if it behaves as is expected. Two potential methods to contribute to this analysis are standard questions and sensitivity analysis.

3.1 Standard Questions

Standard questions comprise a list of questions the modeller (and potentially an independent expert) should ask about the construction, operation and output of the model. They help to identify uncertainty in model components, unexpected behaviour and areas where improvement is required. (Parker et al, 2002 and Risbey et al, 1996) A list of standard questions is provided in

Table 1.

Table 1: Qualitative questions for model evaluation.

No	Question
9.1	How reliable is the input data? What are the uncertainties in the input data? (e.g. measurement error, sampling rate) How does this affect the model?
9.2	Is the model behaving as expected? How is the model behaviour affected by any assumptions required for the development of the model?
9.3	Does the model structure reflect the system i.e. Is the model structure plausible?
9.4	Is the model over-fitted?
9.5	Is the model flexible/transparent?
9.6	Have alternative model structures/types been tested? Why was the current model structure selected?
9.7	Does the model meet its specified purpose?
9.8	How realistic and optimal are selected parameter values?

4 SENSITIVITY ANALYSIS

Sensitivity analysis explores how a change in parameter values effect the overall change in the output of the model. This can be completed using simple sensitivity analysis, where only one parameter is changed or more complex arrangements that explore the relationships between multiple parameters. This analysis allows the opportunity for more extensive face validation of a model, where the behaviour of each parameter can be compared to the expected behaviour (Saltelli et al, 2000).

Global/sampling methods sample each parameter over their entire distribution to examine the sensitivity of the model. A simple approach to this could simply be random sampling based on the expected distributions of the parameters (Monte Carlo analysis) and using simple visualisations or regression/correlation analysis to examine the model. It is possible to use more complicated methods which use transformation functions to sample the parameter space. For example FAST which uses a Fourier transform function (Cukier et al, 1978) or Sobol' which uses a dimensionality decomposition (Sobol', 1993). More complicated search algorithms that optimize the parameter sample (e.g. genetic algorithms) can also be used. For large complex models this analysis can still be computationally expensive and one option is to utilise transformation functions to represent the model with less complexity. An example of this is high dimensional model representation (e.g. Ziehn and Tomlin, 2008) that represents the mapping between input and output with polynomials of different orders.

Algebraic sensitivity analysis takes a different approach by directly examining the equations of the model. For each operation the sensitivity is calculated and then combined algebraically for the model operations. It is completed by considering finite proportional changes to the input and deriving how it changes the output function, applying simplifications where appropriate. A thorough introduction to the method including derivation of sensitivities for basic operations is provided in Norton (2008). This method has many potential benefits including extra insight into the observed sensitivity behaviour of the model. For larger models calculation and derivation may become more difficult. But complexity could be reduced by simplifications of the analysis method or by performing the analysis on individual components independently.

5 COMBINING QUALITATIVE WITH QUANTITATIVE TESTING

A thorough testing procedure will include both qualitative and quantitative evaluation. When this occurs it is necessary to consider systematically both the qualitative and quantitative components. In some modelling communities this has lead to the development of systematic protocols that allow for the consideration of both factors, The Good

Modelling Practice Handbook (STOWA/RIZA, 1999) for deterministic, numerical models and guidelines for groundwater modelling by the Murray-Darling Basin Commission (2000) are two examples of checklists developed to evaluate models systematically.

Another approach is to assign numerical values to each question allowing the models to be rated either numerically or graphically. One system that uses this approach is the Numerical Unit Spread Assessment Pedigree (NUSAP) system. This system combines derived numerical metrics (including some form of error calculation and spread calculation) with more qualitative approaches used to assess the performance of the model and the process used to generate the model. The results from these multi-criteria tests are combined onto a single kite-diagram allowing easy comparison of various models' performance.

6 APPLICATION RECOMMENDATIONS

As part of their 10 step modelling procedure Jakeman et al (2006) define a set of minimum standards which models should include (but not be limited to). These standards are repeated below:

1. Clear statement of the objectives and clients of the modelling exercise;
2. Documentation of the nature (identity, provenance, quantity and quality) of the data used to drive , identify and test the model;
3. A strong rationale for the choice of model families and features (encompassing alternatives);
4. Justification of the methods and criteria employed in calibration;
5. As thorough analysis and testing of model performance as resources allow and the application demands;
6. A resultant statement of model utility, assumptions, accuracy, limitations, and the need and potential for improvement; and quite obviously but importantly;
7. Fully adequate reporting of all of the above, sufficient to allow informed criticism.

These standards are all applicable to the performance evaluation process including the selection and application of the performance criteria. Each modelling task completed will likely have unique goals and challenges, which means there is no ideal standard technique applicable for all models. However, despite their differences it is possible to suggest a general procedure that would be beneficial to many models. The procedure suggested is summarised in the following four steps.

Step 1: Identification of the model's purpose

The most important step of the procedure is the initial step, it is necessary to have a clear idea of the modelling purpose. This means having a clear idea of what events are being modelled and what will constitute a 'good' model. Having a clear idea of the model's purpose allows easy selection of error metrics.

Step 2: Identification of data characteristics

The second step involves an analysis of data which is used to test the model. This involves determining whether there is any data available and how much of this data is required for the development and calibration of the model (generally more calibration data is required for models of greater complexity) At this point it should also be determined whether there is enough data and computing resources/time to consider multiple calibration and testing periods.

The data can then be analysed. For the initial analysis a graphical procedure is suggested to detect the general behaviour of the data to be modelled. For time series data, an auto-correlation procedure will detect any periodicity in the data, while calculating the empirical distribution function will give a better impression of the magnitude of events. It may be necessary to examine a time domain plot of events to detect during what period events and outliers occur. At the completion of these tests there will be a clearer understanding of the data in the system and period/s for calibration and testing can be selected with confidence.

Step 3: Graphical performance analysis

The third step entails a graphic analysis to judge the performance of the model. From this step there are two main goals: the detection of likely under- or non-modelled behaviour and gaining an overview of the overall performance of the model. The residual plot, QQ plot and a cross-correlation between the input data and residuals are all capable of indicating when a model is not completely representing a system's behaviour. These results can be used to judge the model's performance or to help refine the model before the rest of the evaluation is completed.

Step 4: Select basic performance criteria

It is necessary to select performance criteria to evaluate the model. Root mean square error (RMSE) or Nash-Sutcliffe efficiency (R^2) are ideal candidates for an initial metric as their widespread usage will benefit in communication of the performance of the model. A thorough understanding of the selected metric is, however, necessary. In particular any weaknesses of a metric for a particular purpose must be addressed. Even in the initial model valuation multiple metrics should be considered. Metrics should be paired which help to overcome the error of individual metrics. For example R^2, which can suffer from a significant offset error should be paired with bias. RMSE (or again R^2) can be paired with a selected data transformation to reduce the effect large events have on the evaluation.

Step 5: Consideration of advanced methods

Once an analysis has been completed using the basic performance criteria it is possible to consider how complete the current evaluation has been. The simple metrics are judged against the knowledge gained from the graphical analysis in step 3, and how well the current evaluation differentiates between multiple models and expert knowledge. Depending on the problems that are identified there are many possible advanced methods that can be considered (Table 2).

Table 2: Selection of Advanced Methods

Problem Identified	Potential Solutions
Changes in model divergence overtime not captured by current metrics	Need a windowed metric, more advanced wavelet analysis, or a metric that is able to allow for the changes in uncertainty (assuming uncertainty is driving the divergence)
Objective functions not effectively differentiating between models	Modify objective function based on sensitivity analysis, or use multi-criteria methods, consider application of Pareto methods in case methods are actually the same
Significant difference between calibration and testing model performance	Period of calibration may not be well chosen, perform sensitivity analysis to determine which parameters are causing trouble, DYNIA for periods parameters are active. Try different/multiple calibration periods.
Significant divergence in low/high magnitude events not captured by metrics	Use data transformations to highlight the differences, metrics that allow for the divergence (e.g. HMLE). Consider multi-resolution methods

REFERENCES

Akaike, H., A new look at the statistical model identification, *Automatic Control, IEEE Transactions 19* (6), 716-723, 1974.

Boyle, D.P., H.V. Gupta, and S. Sorooshian, Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resources Research 36* (12), 3663-3674, 2000.

Choi, H., and K. Beven, Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of topmodel within the glue framework, *Journal of Hydrology 332* (3-4), 316-336, 2007.

Chou, C.-M., Applying multi-resolution analysis to differential hydrological grey models with dual series, *Journal of Hydrology 332* (1-2), 174-186, 2007.

Clark, M.P., A.G. Slater, D.E. Rupp, R.A. Woods, J.A. Vrugt, H. V. Gupta, T. Wagener, and L.E. Hay, Framework for Understanding Structural Errors (FUSE): A module framework to diagnose differences between hydrological models, *Water Resources Research, 44*, W00B02, doi :10.1029/2007WR006735.

Congalton, R.G. A review of assessing the accuracy of classifications of remotely sensed data, *Remote Sensing of Environment, 37*, 35-46, 1991.

Croke, B.F.W. The role of uncertainty in design of objective functions. In Oxley, L. and Kulasiri, D. (eds) MODSIM 2007 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2007, pp. 2541-2547. ISBN: 978-0-9758400-4-7, 2007. http://www.mssanz.org.au/MODSIM07/papers/45_s40/TheRoleOfs40_Croke_.pdf

Croke, B.F.W. Representing uncertainty in objective functions: extension to include the influence of serial correlation. In Anderssen, R.S., R.D. Braddock and L.T.H. Newham (eds) 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, pp. 3372-3378. ISBN: 978-0-9758400-7-8, 2009. http://www.mssanz.org.au/modsim09/I7/croke.pdf

Cukier, R.I., H.B. Levine, and K.E. Shuler, Nonlinear Sensitivity Analysis of Multiparameter Model Systems, *Journal of Computational Physics, 26*, 1-42, 1978.

Dawson, C., R. Abrahart, and L. See, Hydrotest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, *Environmental Modelling & Software 22* (7), 1034-1052, 2007.

Fox, D.G., Judging air quality model performance, *Bulletin of the American Meteorological Society, 62* (5), 599-609, 1981.

Gupta, H.V., S. Sorooshian, and P.O. Yapo, Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research 34* (4), 751-764, 1997.

Gupta, H.V., T. Wagener, and Y. Liu, Reconciling theory with observations: elements of a diagnostic approach to model evaluation, *Hydrological Processes, 22*, 3802-3813, 2008.

Hannachi, A., I.T. Jolliffe, and D.B. Stephenson, Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology 27* (9), 1119-1152, 2007.

Härdle, W., J. Horowitz, and J.P. Kreiss, Bootstrap methods for time series, *International Statistical Review/Revue Internationale de Statistique 71* (2), 435-459, 2003.

Jakeman, A.J., and G.M. Hornberger, How much complexity is warranted in a rainfall-runoff model? *Water Resources Research 29* (8), 2637-2649, 1993.

Jakeman, A.J., R.A. Letcher, and J.P. Norton, Ten iterative steps in development and evaluation of environmental models, *Environmental Modelling & Software 21* (5), 602-614, 2006.

Kelmes, V., Operational Testing of Hydrological Simulation Models, *Hydrological Sciences Journal, 31* (1), 13-24, 1986.

Kleijnen, J.P.C., B. Bettonvil, and W. Van Groenendaal, Validation of trace-driven simulation models: a novel regression test, *Manage. Sci. 44* (6), 812-819, 1998.

Kohavi, R., A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, August 20-25 1995, Montreal Canada, Vol. 2. pp. 1137-1143, 1995.

Lane, S.N., Assessment of rainfall-runoff models based upon wavelet analysis, *Hydrological Processes 21* (5), 586-607, 2007.

Moriasi, D. N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, and T.L. Veith, Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Transactions of the ASABE 50* (3), 885-900, 2007.

Murray-Darling Basin Commission, Groundwater Flow Modelling Guideline. Murray-Darling Basin Commission, Canberra. Project no. 125, 2000.

Norton, J., August Algebraic sensitivity analysis of environmental models, *Environmental Modelling & Software 23* (8), 963-972, 2008.

Oreskes, N., K. Shrader-Frechette, and K. Belitz, Verification, validation, and confirmation of numerical models in the earth sciences, *Science 263* (5147), 641-646, 1994.

Parker, P., R. Letcher, A. Jakeman, M. Beck, G. Harris, R. Argent, M. Hare, C. Pahl-Wostl, A. Voinov, and M. Janssen, Progress in integrated assessment and modelling, *Environmental Modelling & Software 17* (3), 209-217, 2002.

Pontius, R., Useful techniques of validation for spatially explicit land-change models, *Ecological Modelling 179* (4), 445-461, 2004.

Risbey, J., M. Kandlikar, and A. Patwardhan, November Assessing integrated assessments, *Climatic Change 34* (3), 369-395, 1996.

Saltelli, A., K. Chan, and E.M. Scott, Sensitivity Analysis, Wiley Series in Probability and Statistics, Wiley. 2000.

Schwarz, G., Estimating the dimension of a model, *The Annals of Statistics 6* (2), 461-464, 1978.

Sobol', I. M., Sensitivity Estimates for Nonlinear Mathematical Models, *Mathematical Modelling & Computational Experiment, 1*, 407-414. 1993,

Sorooshian, S., and D.A. Dracup, Stochastic Parameter Estimation Procedures for Hydrologic Rainfall-Runoff Models: Correlated and Heteroscedastic Error Cases, *Water Resources Research, 16* (2), 430-442, 1980,

STOWA/RIZA, 1999. Smooth Modelling in Water Management, Good Modelling Practice Handbook. STOWA Report 99-05. Dutch Department of Public Works, Institute for Inland Water Management and Waste Water Treatment, ISBN 90-5773-056-1. Report 99.036.

Taylor, C.J., D.J. Pedregal, P.C. Young, and W. Tych, Environmental time series analysis and forecasting with the Captain toolbox, *Environmental Modelling and Software, 22* (6), 797-814, 2007.

van der Sluijs, J.P., M. Craye, S. Funtowicz, P. Kloprogge, J. Ravetz, and J. Risbey, Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System', *Risk Analysis* 25(2):481-492, 2005.

Wagener, T., D. Boyle, M. Lees, H. Wheater, H. Gupta, and S. Sorooshian, A framework for development and application of hydrological models. *Hydrology and Earth System Sciences, 5*(1), 13-26, 2001.

Wagener, T., N. Mcintyre, M.J. Lees, H.S. Wheater, and H.V. Gupta, Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, *Hydrological Processes 17* (2), 455-476, 2003.

Wealands, S., R. Grayson, and J. Walker, Quantitative comparison of spatial fields for hydrological model assessment - some promising approaches, *Advances in Water Resources 28* (1), 15-32, 2005.

Yapo, P., H. Gupta, and S. Sorooshian, Multi-objective global optimization for hydrologic models, *Journal of Hydrology 204* (1-4), 83-97, 1998.

Young, P.C. Data-based mechanistic modelling and validation of rainfall-flow processes. In M. G. Anderson and P.D. Bates (Eds.), Model Validation: Perspectives in Hydrological Science. Chichester: J. Wiley, 117-161, 2001.

Young, P.C., A.J. Jakeman, and R.E. McMurtrie, An instrumental variable method for model order identification, *Automatic, 16*, 281-94, 1980.

Ziehn, T., and A.S. Tomlin, Gui–hdmr – a software tool for global sensitivity analysis of complex models. *Environmental Modelling & Software 24* (7), 775-785, 2008.