Environmental Modelling & Software 40 (2013) 1–20

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Position paper

Characterising performance of environmental models☆

Neil D. Bennett [a], Barry F.W. Croke [a], Giorgio Guariso [b], Joseph H.A. Guillaume [a], Serena H. Hamilton [a], Anthony J. Jakeman [a,*], Stefano Marsili-Libelli [c], Lachlan T.H. Newham [a], John P. Norton [a], Charles Perrin [d], Suzanne A. Pierce [e], Barbara Robson [f], Ralf Seppelt [g], Alexey A. Voinov [h], Brian D. Fath [i,j], Vazken Andreassian [d]

[a] Fenner School of Environment and Society, National Centre for Groundwater Research and Training, The Australian National University, Australia
[b] Politecnico di Milano, Italy
[c] Department of Systems and Computers, University of Florence, Italy
[d] IRSTEA, France
[e] Center for International Energy and Environmental Policy, Jackson School of Geosciences, The University of Texas at Austin, USA
[f] CSIRO Land and Water, Australia
[g] UFZ – Helmholtz Centre for Environment Research, Department of Computational Landscape Ecology, Leipzig, Germany
[h] ITC, Twente University, The Netherlands
[i] Department of Biological Sciences, Towson University, USA
[j] Advanced Systems Analysis Program, International Institute for Applied Systems Analysis, Austria

ARTICLE INFO

Article history:
Received 17 July 2012
Received in revised form
25 September 2012
Accepted 25 September 2012
Available online 6 November 2012

Keywords:
Model development
Model evaluation
Performance indicators
Model testing
Sensitivity analysis

ABSTRACT

In order to use environmental models effectively for management and decision-making, it is vital to establish an appropriate level of confidence in their performance. This paper reviews techniques available across various fields for characterising the performance of environmental models with focus on numerical, graphical and qualitative methods. General classes of direct value comparison, coupling real and modelled values, preserving data patterns, indirect metrics based on parameter values, and data transformations are discussed. In practice environmental modelling requires the use and implementation of workflows that combine several methods, tailored to the model purpose and dependent upon the data and information available. A five-step procedure for performance evaluation of models is suggested, with the key elements including: (i) (re)assessment of the model's aim, scale and scope; (ii) characterisation of the data for calibration and testing; (iii) visual and other analysis to detect under- or non-modelled behaviour and to gain an overview of overall performance; (iv) selection of basic performance criteria; and (v) consideration of more advanced methods to handle problems such as systematic divergence between modelled and observed values.

1. Introduction

Quantitative environmental models are extensively used in research, management and decision-making. Establishing our confidence in the outputs of such models is crucial in justifying their continuing use while also recognizing limitations. The question of evaluating a model's performance relative to our understanding and observations of the system has resulted in many different approaches and much debate on the identification of a *most* appropriate technique (Alexandrov et al., 2011; McIntosh et al., 2011). One reason for continued debate is that performance measurement is intrinsically case-dependent. In particular, the manner in which performance is characterised depends on the field of application, characteristics of the model, data, information and knowledge that we have at our disposal, and the specific objectives of the modelling exercise (Jakeman et al., 2006; Matthews et al., 2011).

Modelling is used across many environmental fields: hydrology, air pollution, ecology, hazard assessment, and climate dynamics, to name a few. In each of these fields, many different types of models are available, each incorporating a range of characteristics to measure and represent the natural system behaviours.

☆ Position papers aim to synthesise some key aspect of the knowledge platform for environmental modelling and software issues. The review process is twofold – a normal external review process followed by extensive review by EMS Board members. See the Editorial in Volume 21 (2006).

* Corresponding author.
E-mail address: tony.jakeman@anu.edu.au (A.J. Jakeman).

1364-8152/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2012.09.011

Environmental models for management typically consist of multiple interacting components with errors that do not exhibit predictable properties. This makes the *traditional* hypothesis-testing associated with statistical modelling less suitable, at least on its own, because of the strong assumptions generally required, and the difficulty (sometimes impossibility) of testing hypotheses separately. Additionally if a single performance criterion is used, it generally measures only specific aspects of a model's performance, which may lead to counterproductive results such as favouring models that do not reproduce important features of a system (e.g., Krause et al., 2005; Hejazi and Moglen, 2008). Consequently, systems of metrics focussing on several aspects may be needed for a comprehensive evaluation of models, as advocated e.g., by Gupta et al. (2012).

It is generally accepted that the appropriate form of a model will depend on its specific objectives (Jakeman et al., 2006), which often fall in the broad categories of improved understanding of natural processes or response to management questions. The appropriate type of performance evaluation clearly depends on the model objectives as well. Additionally, there may be several views as to the purpose of a model, and multiple performance approaches may have to be used simultaneously to meet the multi-objective requirements for a given problem. In the end, the modeller must be confident that a model will fulfil its purpose, and that a 'better' model could not have been selected given the available resources. These decisions are a complex mixture of objectively identified criteria and subjective judgements that represent essential steps in the cyclic process of model development and adoption. In addition, the end-users of a model must also be satisfied, and may not be comfortable using the same performance measures as the expert modeller (e.g., Miles et al., 2000). It is clear that in this context, a modeller must be eclectic in choice of methods for characterising the performance of models.

Regardless, assessing model performance with quantitative tools is found to be useful, indeed most often necessary, and it is important that the modeller be aware of available tools. Quantitative tools allow comparison of models, point out where models differ from one another, and provide some measure of objectivity in establishing the credibility and limitations of a model. Quantitative testing involves the calculation of suitable numerical metrics to characterise model performance. Calculating a metric value provides a single common point of comparison between models and offers great benefits in terms of automation, for example automatic calibration and selection of models. The use of metric values also minimises potential inconsistencies arising from human judgement. Because of the expert knowledge often required to use these tools, the methods discussed in this paper are intended for use primarily by modellers, but they may also be useful to inform end-users or stakeholders about aspects of model performance.

This paper reviews methods for quantitatively characterising model performance, identifying key features so that modellers can make an informed choice suitable for their situation. A classification is used that cuts across a variety of fields. Methods with different names or developed for different applications are sometimes more similar than they at first appear. Studies in one domain can take advantage of developments in others. Although the primary applications under consideration are environmental, methods developed in other fields are also included in this review. We assume that a model is available, along with data representing observations from a real system, that preferably have not been used at any stage during model development and that can be compared with the model output. This dataset should be representative of the model aims; for instance it should contain flood episodes or pollution peaks if the model is to be used in such circumstances.

The following section provides a brief view of how characterisation of model performance fits into the broader literature on the modelling process. Section 3 reviews selection of a so-called 'validation' dataset. In Section 4, quantitative methods for characterising performance are summarized, within the broad categories of direct value comparison, coupling real and modelled values, preserving data patterns, indirect metrics based on parameter values, and data transformations. Section 5 discusses how qualitative and subjective considerations enter into adoption of the model in combination with quantitative methods. Section 6 presents an approach to selecting performance criteria for environmental modelling. Note that a shorter and less comprehensive version of this paper was published as Bennett et al. (2010).

2. Performance characterisation in context

With characterisation of model performance being a core part of model development and testing, there is naturally a substantial body of related work. This section presents some key links between similar methods that have developed separately in different fields. In many of the fields of environmental modelling, methods and criteria to judge the performance of models have been considered in the context of model development. Examples include work completed for hydrological models (Krause et al., 2005; Jakeman et al., 2006; Moriasi et al., 2007; Reusser et al., 2009), ecological models (Rykiel, 1996) and air quality models (Fox, 1981; Thunis et al., 2012). The history of methods to characterise performance dates back at least a few decades (see, for instance Fox, 1981; Willmott, 1981) and makes use of artificial intelligence (Liu et al., 2005) and statistical models (Kleijnen, 1999), while efforts to standardise them extensively and include them in a coherent model-development chain are more recent. For instance, in hydrology, previous work has been completed on general modelling frameworks that consider performance criteria as part of the iterative modelling process (Jakeman et al., 2006; Refsgaard et al., 2005; Wagener et al., 2001). Stow et al. (2009) present a summary of metrics that have been used for skill assessment of coupled biological and physical models of marine systems. Studies have also focused explicitly on performance criteria, such as Moriasi et al. (2007) and Dawson et al. (2007, 2010), who produced guidelines for systematic model evaluation, including a list of recommended evaluation techniques and performance metrics. Beck (2006) provides a survey of key issues related to performance evaluation. And Matott et al. (2009) reviewed model evaluation concepts in the context of integrated environmental models and discussed several relevant software-based tools.

Some official documents on model evaluation have also been produced by governing agencies. Among these, some are particularly detailed, e.g., "Guidance on the use of models for the European Air Quality Directive" (FAIRMODE, 2010), the "Guidance for Quality Assurance Project Plans for Modelling" (USEPA, 2002) and the "Guidance on the Development, Evaluation, and Application of Environmental Models" (USEPA, 2009). This paper, by contrast, focuses on graphical and numerical methods to characterise model performance. The domain-specific reviews are synthesised for a broader audience. Use of these methods within the modelling process is only briefly discussed, and the reader is directed to other references for more detail. Finally, a philosophical debate has aimed to differentiate verification from validation (Jakeman et al., 2006; Oreskes et al., 1994; Refsgaard and Henriksen, 2004). We, however, focus on summarizing methods to characterise performance of environmental models, whether these methods and criteria are used for verification, validation or calibration instead of continuing this debate.

While this paper focuses mainly on model evaluation by comparison to data, the reader should be aware of cases where this is not possible due to the nature of the analysis or the stage of modelling. For example, such evaluation techniques are not possible for qualitative conceptual models built in participatory modelling frameworks to establish a common understanding of the system. The methods presented may not be sufficient for complex situations, for example where the identification of feedbacks and relationships between different environmental processes across spatial and temporal scales is of specific interest. Untangling feedbacks, especially in highly complex and spatially interrelated systems, remains a big challenge (Seppelt et al., 2009). Also, performance in future conditions cannot be directly assessed as available data may not be representative; this is particularly the case where the model includes an intervention that will change the behaviour of the system.

Data-based performance measures are only a small part of quality assurance (QA) in modelling (Refsgaard et al., 2005). Guidelines and tools such as those produced by RIVM/MNP (van der Sluijs et al., 2004), the HarmoniQuA project (Henriksen et al., 2009) and Matott et al. (2009) generally cover the qualitative definition of the model requirements and stakeholder engagement, in addition to the quantitative measures discussed in this paper. A consistent procedure for characterising performance over the entire model life cycle can facilitate QA by providing clarity, and hence increased knowledge and confidence in developing and selecting the most appropriate model to suit particular goals. This can also benefit future model reuse and application.

The paper will not discuss formal methods for sensitivity or uncertainty analysis, though these aspects are recognised as central in the modelling process. Sensitivity analysis assesses how variations in input parameters, model parameters or boundary conditions affect the model output. For more information on sensitivity analysis, see Saltelli et al. (2000), Norton (2008), Frey and Patil (2002), Saltelli and Annoni (2010), Shahsavani and Grimvall (2011), Makler-Pick et al. (2011), Yang (2011), Nossent et al. (2011), Ratto et al. (2012) and Ravalico et al. (2010). Uncertainty analysis as generally understood is concerned with establishing bounds around point predictions, from either deterministic analysis based on model-output error bounds or probabilistic analysis yielding confidence intervals. While this is often a useful measure of model performance, related methods have already been reviewed elsewhere (e.g., Beven, 2008; Keesman et al., 2011; Vrugt et al., 2009). A distinct set of measures can be used to evaluate interval or probabilistic predictions. While not covered in this paper, the reader is referred to a number of useful reviews (Laio and Tamea, 2007; Gneiting and Raftery, 2007; Murphy and Winkler, 1987; Christoffersen, 1998; Boucher et al., 2009).

We stress the distinction between characterising performance and adoption or rejection of the model by the modeller and end-user. This paper discusses objective measures to assist the first, though the application of measures and their interpretation have a subjective element. The second is completely subjective and may include factors not strictly connected to the measured performance of the model such as costs, simplicity, applicability, intelligibility by the user(s), or how the model is presented. The adoption of a model (or confidence or trust in its ability to provide meaningful insights) is the output of a subjective, typically qualitative and often hidden, psychological or sometimes political process. The implications of this distinction are discussed further in Section 5. This choice can, however, be made more objective by defining benchmarks that can help evaluating models in a comparative way, as advocated for example by Seibert (2001), Matott et al. (2012) and Perrin et al. (2006a, 2006b).

3. Data for characterising performance

The most important component of quantitative testing is the use of observational data for comparison. However, some of the data must be used in the development and calibration (if required) of the model. This necessitates the division of available data to permit development, calibration and performance evaluation. Common methods for this division are presented in Table 1 and include cross-validation and bootstrapping. In cross-validation, the data are split into separate groups for development and testing, whereas bootstrapping involves repeated random sampling with replacement of the original measurements. In spatial modelling, the concept of data division can include separation using both temporal and spatial subdivision. Instead of segmenting a single set of time domain data, a set of spatial data can be removed for testing; for example, distributed hydrological models can be tested by leaving out one gauge at a time. In using such techniques, it is important to consider the degree to which the verification or validation data are independent of the calibration data – for instance, leaving out one gauge will not allow a meaningful test when gauges are closely spaced along a single stretch of river.

Table 1
Examples of data-division methods for model testing.

Type	Name	Description
No independent data testing	Re-substitution	The same data are used for development and testing. Therefore, the performance evaluation will be the same as the calibration evaluation and model performance is likely to be overestimated because the model has been 'tuned' to the results. This approach is the least rigorous, and should be avoided.
Cross-validation	Hold out method	The data are split into two groups, one for development and one for testing. The size and position of the group splitting will affect both the performance of the model and the accuracy of the testing (Kohavi, 1995).
	K-fold partitioning	Data split into k sets, one set is used for training, the remaining $k-1$ sets used for testing. The hold out method can then be repeated k times allowing all results to be averaged (Kohavi, 1995).
	Leave one out (LOOCV)	Here $n-1$ data points are used for model development and only one point is used for validation. This is repeated for all data points, each one in turn being left out (Kohavi, 1995).
Bootstrapping		Bootstrapping involves random re-sampling with replacement of the original measurements (input and output). This can be repeated multiple times to estimate the error (bootstrap) distribution of the model. However, it is essential to have identically and independently distributed (i.i.d.) residuals, so some estimated transformation of model errors is required. One method suggested for time series (or non i.i.d.) data is to use a blocks method where the blocks are randomly re-sampled with replacement. Kleijnen (1999) suggests a method using replicated runs.

Selecting a "representative" set of data that sufficiently captures the key patterns in the data involves many considerations, for example, the sampling method, the heterogeneity of the measure, and sources of sampling errors; for further discussion see Cochran (1977), Gy (1998) and Nocerino et al. (2005). One may want to test the ability of the model to generalise, i.e. its ability to predict sufficiently accurately for a given purpose, even in conditions that were not observed during model development. Ideally, model results are compared to unseen data from the past that capture the conditions that the model will be used to predict. Models are often used to simulate extreme conditions; for such cases the testing data should include the relevant conditions, for example, data that covers a particularly warm or wet period. Models are more likely to fail against such testing data if calibrated primarily over less extreme periods, but will provide greater confidence if successful (e.g., Robson and Hamilton, 2004; Andréassian et al., 2009; Coron et al., 2012; Seiller et al., 2012). In other cases, however, the data available do not sufficiently cover the prediction conditions, so confidence in the model performance should instead be built by looking at components of the overall model or generating suitable artificial "data" sequences. Such artificial data sequences may be generated with a more sophisticated (and more expensive or difficult to use) model in which confidence has already been established through other means (e.g., Raick et al., 2006; Littlewood and Croke, 2008); for example, a flow-routing model may be tested against output from a detailed computational fluid dynamics model. Whatever the case, the decision of what is representative is subjective and strongly linked to the final aim of the model.

Metrics typically calculate a single value for the whole dataset, which can disguise significant divergent behaviour over the time intervals or spatial fields (see for example Berthet et al., 2010; Moussa, 2010). To prevent this, evaluation can instead be on a local basis (e.g., pixel to pixel or event to event). Data may be partitioned *a priori*, using some external information about, for instance, the season or the specific spatial domain, or they can be separated into low, medium, and high events (Yilmaz et al., 2008) or, in hydrology, into rising-hydrograph and falling-hydrograph components, each with its own performance criteria. A more complex partitioning scheme in hydrology suggested by Boyle et al. (2000) involves splitting the data into components driven by rainfall, interflow and baseflow components. As another alternative, Choi and Beven (2007) use a multi-period framework, where a moving window of thirty days is used to classify periods into dry, wetting, wet and drying climate states.

Automatic grouping methods may be used to select approximately homogeneous regions or regions of interest from the original data to assess model performance under different conditions. Wealands et al. (2005) studied the methods used in image processing, landscape ecology and content-based image retrieval among others, and recommended several methods as potentially useful for application to spatial hydrological models. They recommend a clustering method that starts with all the pixels as separate regions, which are then merged wherever the gradient in the variable of interest is below a threshold. Merging is continued in an iterative process until no neighbouring regions satisfy the criteria for merging. Further control of the regions can be achieved by imposing additional criteria, for example criteria on the desired shape, size and roundness of the regions (Wealands et al., 2005).

Evidently any of these methods can be used in both the time and spatial domains, but sometimes it is necessary to consider both spatial and temporal performance, which may require the selection of a 4-dimensional dataset to allow a combination of spatial mapping of temporally averaged metrics and time-series representation of spatially averaged metrics (e.g., Robson et al., 2010).

It is important to realise that, just as the model output is not the same as the true state of the environmental system, neither is the observational dataset. Rather, the observational data provide (imperfect) evidence regarding the true state of the system. Measurement errors, spatial and temporal heterogeneity at scales below the resolution of measurements, and the distinction between what is measured (e.g., chlorophyll *a* or even fluorescence) and what is modelled (e.g., phytoplankton concentrations converted to approximate chlorophyll concentration for comparison purposes) all contribute to the error and uncertainty in the degree to which the observational data reflect reality. What this means for assessing model performance is that not only is it almost impossible to achieve an exact match between model and data, it may not even be desirable. Bayesian techniques for parameter estimation and data assimilation can allow for this by treating both the model and the data as priors with different degrees of uncertainty (Poole and Raftery, 2000). When using non-Bayesian approaches, a modeller should aim to understand the degree of error inherent in the data before setting performance criteria for the model.

4. Methods for measuring quantitative performance

Quantitative testing methods can be classified in many ways. We use a convenient grouping based on common characteristics. Direct value comparison methods directly compare model output to observed data as a whole (4.1). These contrast with methods that combine individual observed and modelled values in some way (4.2). Within this category, values can be compared point-by-point concurrently (4.2.1), by calculating the residual error (4.2.2), or by transforming the error in some way (4.2.3). The relationship between points is considered in methods that preserve the data pattern (4.3). Two completely different approaches measure performance according to parameter values (4.4), and by transformation of the data to a different domain (4.5), a key example being the use of Fourier transforms.

4.1. Direct value comparison

The purpose of direct value comparison methods is to test whether the model output y (the elements of an array of dimension 1–4 depending on the number of independent variables: time and/or one or more spatial coordinates) shows similar characteristics as a whole to the set of comparison data $\hat{y}$ (having the same dimensions but not necessarily the same granularity). The simplest direct comparison methods are standard summary statistics of both y and $\hat{y}$ shown in Table 2. Clearly, one would like the summary statistics computed on y to be very close in value to those computed on $\hat{y}$. Common statistical metrics that can be used for the direct comparison of models and test data include the mean, mode, median and range of the data. The variance, a measure of the spread of the data, is often computed. Higher-order moments, such as kurtosis and skew, could also potentially be used for comparison. Note that in comparing model and observation, statistical properties may be complicated by different support scales: for instance, if the model resolution means that it is averaging over a 1 km grid and a 1 day time-step, whereas observations depend on instantaneous 100 mL water samples, a lower variance might legitimately be expected in the model results, without invalidating the model.

A related method involves comparison of empirical distribution functions, plotted as continuous functions or arranged as histograms. Often the cumulative distributions of the modelled output and the observed data are estimated according to Equation (2.9) in Table 2. The two distributions can then be directly compared. When time is the only independent variable, the cumulative distribution can be interpreted as a "duration curve" showing the fraction of

Table 2
Details of methods for direct comparison of models, where n is the number of observations of y and y_i is the ith observation.

ID	Name	Formula	Notes
2.1	Comparison of scatter plots	~	Look for curvature and dispersion of plots (Figs. 2 and 3).
2.2	Mean	$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$	Calculation of the expected values of modelled and measured data. Need to consider the effect of outliers on each calculation.
2.3	Mode	~	Calculation of most common value in both modelled and measured data.
2.4	Median	~	Unbiased calculation of 'middle' value in both modelled and measured data.
2.5	Range	$\max(y) - \min(y)$	Calculates the maximum spread of data, may be heavily affected by outliers.
2.6	Variance σ^2	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$	Provides a measure of how spread out the data are.
2.7	Skew	$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{3/2}}$	A measure of the asymmetry of the data, skew indicates if the mean of the data is further out than the median. A negative skew (left) has fewer low values and a positive skew (right) has fewer large values.
2.8	Kurtosis	$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{2}} - 3$	Kurtosis is a measure of how peaked the data is. A high kurtosis value indicates that the distribution has a sharp peak with long and fat tails.
2.9	Cumulative distribution	$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le x)$	Empirical distribution of data, compare graphically on normal or logarithmic axis (Fig. 1). Or create a simple error metric.
2.10	Frequency distribution plot, histogram	~	Separate the data into classes, and count the number or percentage of data points in each class.
2.11	Comparison of autocorrelation and cross-correlation plots	~	Graphically compare correlation functions derived from observed and modelled data (or create simple error metrics) to establish whether the model system accurately reflects patterns in the expected direction of change.

time for which the variable exceeds a given value. This interpretation is often useful when the modelled system is going to be used for production (e.g., river flow for water supply, wind speed for a wind generator). Fig. 1 shows the use of a flow–duration curve in hydrological modelling. A logarithmic transformation can be used to highlight the behaviour of the data for smaller events. Yilmaz et al. (2008) present a number of hydrological "signature measures" making use of the flow–duration curve. Simple metrics can be calculated by sampling the curve at specific points of interest. For instance, the duration of time below a certain water level may help in evaluation of model behaviour in low-flow conditions, or the interval above a certain pollution concentration to evaluate performance for dangerous pollution episodes. The integral of the cumulative distribution between two points provides another potential metric. SOMO 35 (sum of daily maximum 8-h values exceeding 35 ppb) and AOT40 (sum of the positive differences between hourly concentration and 40 ppb) are examples of this metric for ozone air pollution.

To generalise from these common examples, many summary measures can be calculated from a given dataset. Comparing them to the same summary measures calculated from the model output provides a (possibly graphical) measure of performance. For example, an autocorrelation or cross-correlation plot of a time series provides information about the relation between points over time within the dataset, whereas box (or box-and-whisker) plots provide a convenient visual summary of several statistical properties of a dataset as they vary over time or space. We expect that the plots from the modelled and observed datasets would be similar, and differences may help identify error in the model or data. Cross-correlations will be discussed again later in the context of residuals. There is an important distinction between the two applications. The methods in this section do not directly compare individual observed and modelled data points. They are therefore suited to evaluate properties and behaviours of the whole dataset (or chosen subsets).

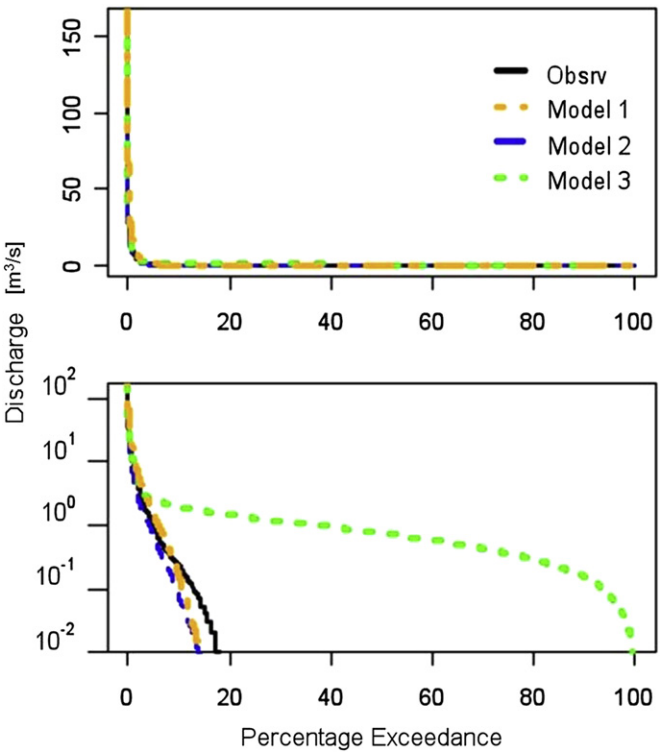

Fig. 1. Empirical cumulative distribution function used for model validation. The flow duration curve is calculated for a rainfall–runoff model, IHACRES (Jakeman et al., 1990). In a standard plot (top) it is difficult to distinguish between results. By using a log transform on the y-axis (bottom) significant divergent behaviour can be observed by Model 3 at low flow levels.

4.2. Coupling real and modelled values

These methods consider the pairs of values y_i and $\hat{y}_i$ for the same point in time or space, i, where y_i is the observed value and $\hat{y}_i$ is the modelled value. These methods often explicitly compute their difference, the residual or error.

4.2.1. Concurrent comparison

The simplest method is the scatter plot (Figs. 2 and 3), where the modelled output values are plotted against the corresponding measured data. An ideal, unbiased model would yield a unity-slope line through the origin; the scatter of the points about the line represents the discrepancy between data and model. Systematic divergence from the line indicates unmodelled behaviour. This graph (with arithmetic or logarithmic scales) is ideal for comparing model performance at low, medium, and high magnitudes, and may well reveal that the model underestimates or overestimates in a certain range if most points lie below or above the line.

The scatter plot can be analysed by computing the statistical properties of the observed data/model data regression line, through an *F*-statistic function of the regression line coefficients (slope and intercept in ID 3.10, Table 3). This function provides a way of checking whether the regression line is statistically similar to the 1:1 line (perfect agreement) (Haefner, 2005). A drawback of this method is that the time-indexing of the data is lost.

An additional test is linear regression analysis on the measured data and model output (Table 3, ID 3.10). When used for hypothesis testing, there are strict requirements for the residuals to be identically and independently (normally) distributed (i.i.n.d). The zero intercept and unit slope of a perfect model can be contrasted with the actual intercept and slope to check if the difference is statistically significant (e.g., with Student's *t*-statistic) or to evaluate the significance of any bias. In hypothesis testing, this significance test is often found to be inadequate, resulting in incorrect acceptance or rejection. However, even if these requirements are not met, it may still be informative for model performance comparison. Kleijnen et al. (1998) proposed a new test where two new time series are created: the difference and the sum of the model and observed values (Table 3, ID 3.12). A regression model is fitted between these new time series; an ideal model has zero intercept and unit slope, since the variables will have equal variances only if their sum and variance are uncorrelated.

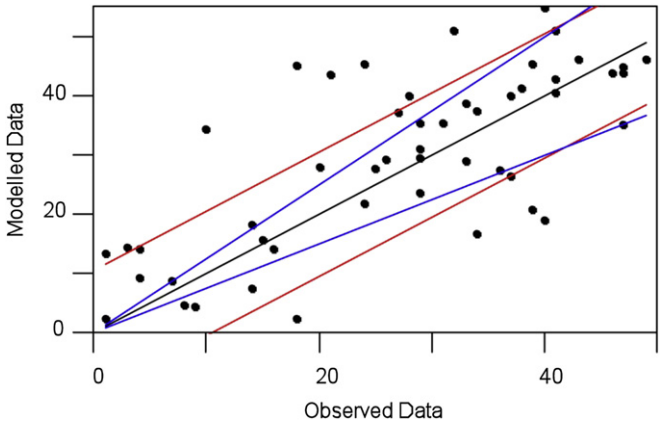

Fig. 2. Scatter plot used for model verification. Modelled and observed data are plotted against each other, residual standard deviation (red) or percentage variance (blue) lines can be plotted to assist in interpretation of the results. Data in this example were from a random number generator. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

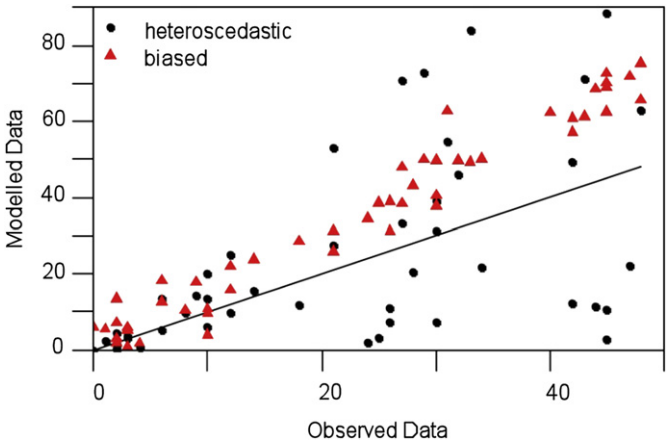

Fig. 3. Scatter plots can reveal underlying behaviour of the model, including bias or non-constant variance. Data in this example were from a random number generator with appropriate relations to introduce bias.

Among the methods coupling real and modelled values, many are based on a table reporting their behaviour in important cases or events, typically passing a specified threshold (commonly referred to as an "alarm" or "trigger" value representing, for instance, a flood level or a high pollution episode). The contingency table (Fig. 4) reports the number of occurrences in which real data and model output were both above the threshold (*hits*), the number in which they were both below (*correct negative*), the number of alarms missed by the model (*misses*) and that of *false alarms*. A perfect model would have data only on the main diagonal. The numbers in the other two cells are already possible metrics for evaluating model under-estimation (high misses) or over-estimation (high false alarms). Many other metrics can be derived from the contingency table (see Ghelli and Ebert, 2008); some are summarized in Table 3. The main purpose of these metrics is to condense into a single number the overall behaviour of the model in terms of critical conditions. If one wants to work on single events, then a quantitative measure of the difference between model and data can also be measured, such as metrics PDIFF and PEP (defined in Table 3, ID 3.8 and 3.9, respectively).

Considering more categories beyond simple occurrence or not of an event can extend this approach. This is so when more than one threshold is defined (e.g., attention and alarm levels) or in spatial problems where points can belong to several different classes. The matrix built in these cases is often termed the "confusion matrix" (Congalton, 1991), comparing the observed and modelled data (Fig. 5) in each category. From the matrix, the percentage of correct identification for one or all categories or other indices in Table 3 can be calculated.

A criticism of these methods is that they do not account for random agreement. A simple metric that tries to account for predictions that are correct by chance is the Kappa statistic (Table 3, ID 3.13). It compares the agreement between the model and observed data against chance agreement (the probability that the model and data would randomly agree). The statistic can confirm if the percentage correct exceeds that obtained by chance. However, the chance percentage can provide misleading results as a low kappa (i.e. 0) could result for a model with good agreement if one category dominates the data. For example, if there are few observations in one category, then the model may completely fail to identify this category while maintaining a high number of correct identifications for the others.

Originally used in image processing, the 'Information' MSE (IMSE) aims at accounting for a subset of data that is more significant, for example, special areas which have higher environmental

Table 3
Details of metrics that compare real and modelled values concurrently.

ID	Name	Formula	Range	Ideal value	Notes
3.1	Accuracy (fraction correct)	$\frac{\text{hits} + \text{correct negatives}}{\text{total}}$	(0, 1)	1	It is heavily influenced by the most common category, usually "no event".
3.2	Bias score (frequency bias)	$\frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}}$	$(0, \infty)$	1	Measures the ratio of the frequency of modelled events to that of observed events. Indicates whether the model has a tendency to underestimate ($BIAS < 1$) or overestimate ($BIAS > 1$).
3.3	Probability of detection (hit rate)	$\frac{\text{hits}}{\text{hits} + \text{misses}}$	(0, 1)	1	Sensitive to hits, but ignores false alarms. Good for rare events.
3.4	False alarm ratio	$\frac{\text{false alarms}}{\text{hits} + \text{false alarms}}$	(0, 1)	0	Sensitive to false alarms, but ignores misses.
3.5	Probability of false detection (false alarm rate)	$\frac{\text{false alarms}}{\text{correct negatives} + \text{false alarms}}$	(0, 1)	0	Sensitive to false alarms, but ignores misses.
3.6	Threat score (critical success index, CSI)	$\frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}}$	(0, 1)	1	Measures the fraction of observed cases that were correctly modelled. It penalizes both misses and false alarms.
3.7	Success index	$\frac{1}{2}\left(\frac{\text{hits}}{\text{hits} + \text{misses}} + \frac{\text{correct negatives}}{\text{observed no}}\right)$	(0, 1)	1	Weights equally the ability of the model to detect correctly occurrences and non-occurrences of events.
3.8	PDIFF Peak Difference	$\max(y_i) - \max(\hat{y}_i)$	–	0	Compares the two largest values from each set, should be restricted to single event comparisons (Dawson et al., 2007).
3.9	PEP Percent Error in Peak	$\frac{\max(y_i) - \max(\hat{y}_i)}{\max(y_i)} * 100$	(0, 100)	0	Percent error in peak is similar to the PDIFF (3.8) calculation except it is divided by the maximum measured value. Only suitable for single events (Dawson et al., 2007).
3.10	*F*-statistic of the regression line	$F_{2,n-2,\alpha} = \frac{na^2 + 2a(b-1)\sum_{i=1}^{n} x_i + (b-1)^2 \sum_{i=1}^{n} x_i^2}{2S_{yx}^2}$			Analyse the coefficients (*a*,*b*) of the data/model output regression line at the α level of confidence.
3.11	Regression analysis (basic)	$y = \beta_0 + \beta_1 \hat{y}$	$(-1, 1)$	1	Perform a simple linear regression to calculate β_0 and β. Ideal values are $\beta_0 = 0$ and $\beta_1 = 1$.
3.12	Regression analysis (novel)	$d_i = y_i - \hat{y}_i$ $u_i = y_i + \hat{y}_i$ $d_i = \gamma_0 + \gamma_1 u_i$	–	–	Perform a simple linear regression to calculate γ_0 and γ_1. Ideal values are $\gamma_0 = 0$ and $\gamma_1 = 0$ (Kleijnen et al., 1998).
3.13	Kappa statistic κ	$\frac{\Pr(a) - \Pr(c)}{1 - \Pr(c)}$	$(-1, 1)$	1	$\Pr(a)$ is the relative agreement and $\Pr(c)$ is the hypothetical probability that both would randomly agree. A value close to zero indicates that the majority of agreement may be due to chance.
3.14	Information Mean Square Error (IMSE)	$\frac{1}{N}\sum_{x,y}(A(x,y)I(x,y) - \hat{A}(x,y)\hat{I}(x,y))^2$ $I(i) = \log_n \frac{1}{P(i)},\ P(i) = \frac{n_i}{N}$	$(0, \infty)$	0	A is the spatial field, I is the information weighting field calculated from a single event grid, P is the probability of occurrence of a given value bin, i is the bin number, n_i is the total number of grid values in the bin and N is the total number of pixels.
3.15	Fuzzy maps	–	–	–	Use of fuzzy relational characteristics to evaluate categorisation models. Different categories and locations are weighted by their relationship.

importance. Tompa et al. (2000) select the weighting of the MSE from an 'event of importance'. Spatial results are split into groups from which weighting according to the probability of the event (pixel value) occurring is calculated. Here, pixels with a low probability of occurrence receive higher weighting (e.g., peak or very low events). The 'event of importance' will significantly affect the outcome and must be chosen with care.

For some categorical models, some categories may be more strongly related than others, representing a smaller error. Similarly a small position error may not be very significant and should not be treated as total disagreement. Fuzzy sets allow the expression of a degree of membership of a category. A common application is the use of fuzzy maps, with special relationship weights defined for locations and categories (Wealands et al., 2005).

Traditionally, the Kappa statistic has been a standard for comparison in spatial models. It has been criticized for inability to track the location of error (Kuhnert et al., 2006), and more recently Pontius and Millones (2010) have pronounced "death to Kappa"

		Observed Events: Yes	Observed Events: No	Total
Modelled Events	Yes	Hits	False Alarms	Modelled Yes
	No	Misses	Correct Negatives	Modelled No
	Total	Observed Yes	Observed No	Total

Fig. 4. Standard structure of a contingency table.

and declared the "birth of Quantity Disagreement and Allocation Disagreement for Accuracy". Kuhnert et al. (2006), and earlier Costanza (1989), also insisted that comparisons based on visual proximity, using a sliding window or expanding window approaches, are more reliable for estimating allocation errors (see Seppelt and Voinov, 2003 for applications).

As a further extension, it may be possible to remove the effect of a well-understood error, to allow errors of different origins to be quantified. Pontius (2004) presents a method for spatially explicit land-change models. Once an initial error calculation has been completed, the maps are adjusted to remove location and magnitude errors.

4.2.2. Key residual methods

By far the most prevalent methods for model evaluation are residual methods, which calculate the difference between observed and modelled data points. The residual plot (Fig. 6) and the QQ plot (Fig. 7) are two simple graphical methods to analyse model residuals.

The residual plot is a plot of residual error as dependent variable and a chosen descriptor variable (e.g., time or location). The plot reveals unmodelled behaviour when there is systematic divergence from zero. For instance, high density of negative values indicates that the model tends to underestimate correct values (in that time or place). If residuals are due to unsystematic measurement error alone, then we may expect them to be normally distributed. The QQ plot (Fig. 7) tests whether or not the distribution of residuals approximates normality. The quantiles of the residuals are plotted against the Gaussian quantiles. Deviations from a straight line indicate the distribution of residuals is skewed towards larger or smaller values and whether it has a relatively 'peaky' or flat distribution.

The statistical significance of the QQ plot derived from a given cumulative distribution can be assessed with the Kolmogorov–Smirnov (KS) or the Lilliefors tests for a given level of confidence. The KS test checks the hypothesis that the two datasets come from the same distribution, whatever it is, whereas the Lilliefors test does the same, but is limited to the Gaussian distribution. This latter test is particularly robust because it does not require one to estimate the null distribution, whereas in the KS test the reference distribution must be provided.

Of the many possible numerical calculations on model residuals, by far the most common are bias and Mean Square Error (MSE). Bias (Table 4, ID 4.3) is simply the mean of the residuals, indicating

		Observed: Category A	Observed: Category B	Observed: Category C
Modelled	Category A	Aa	ab	ac
	Category B	Ba	bb	bc
	Category C	Ca	cb	cc

Fig. 5. Example of confusion matrix where categorical results are tabulated.

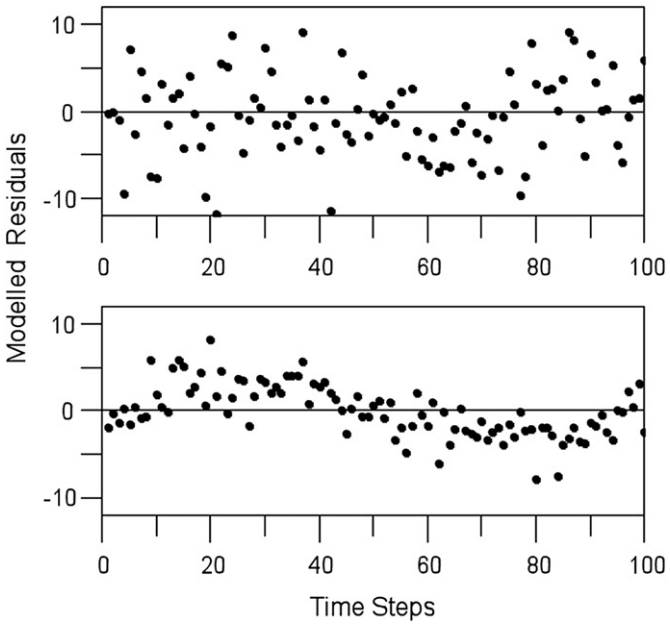

Fig. 6. Residual plot showing the residuals of the model plotted against the descriptor variable (time). A uniform spread of residuals is expected (top), and systematic changes over time indicate unmodelled behaviour (bottom). Data in this example were artificial.

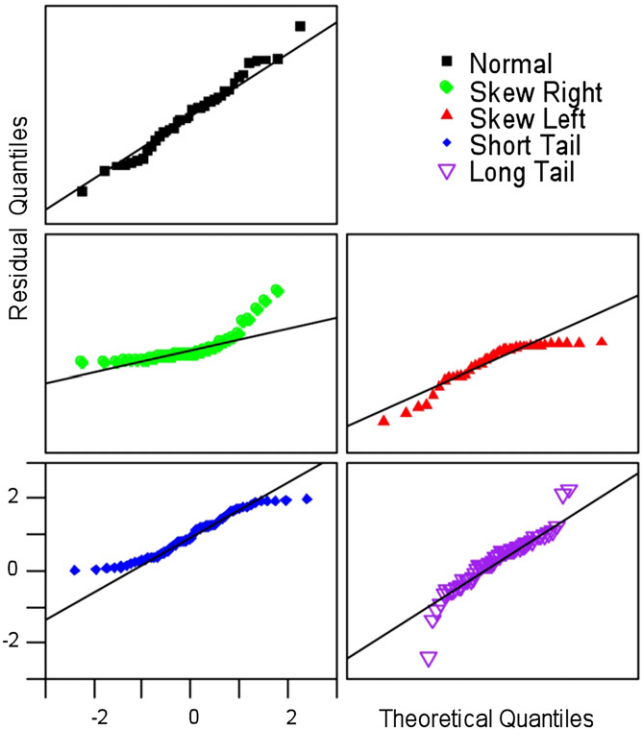

Fig. 7. QQ plot of model residuals. The residuals are compared against the normal distribution and deviation from the line indicates different distribution properties. Data in these examples were from random number generators with the skew and distribution modified to show the different behaviours.

Table 4
Key residual criteria.

ID	Name	Formula	Range	Ideal value	Notes
4.1	Residual plot	~	–	–	Plot residuals against the predictor variable(s), look for curvature or changes in magnitude as the predictor variable changes.
4.2	QQ plot	~	–	–	Plots the inverse distribution (quantile) function of residuals against normal distribution quantile function. Look for curvature and divergence away from the mean diagonal (Fig. 7).
4.3	Bias	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)$	$(-\infty,+\infty)$	0	Calculates the mean error. Result of zero does not necessarily indicate low error due to cancellation.
4.4	Mean Square Error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$(0,\infty)$	0	Calculates a mean error (in data units squared), which is not effected by cancellation. Squaring the data may cause bias towards large events.
4.5	Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$	$(0,\infty)$	0	MSE error (4.4) except result is returned in the same units as model, which is useful for interpretation.
4.6	Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$	$(0,\infty)$	0	Similar to RMSE (4.5) except absolute value is used instead. This reduces the bias towards large events; however, it also produces a non-smooth operator when used in optimisation.
4.7	Absolute Maximum Error (AME)	$\max\lvert y_i-\hat{y}_i\rvert$	–	–	Records the maximum absolute error.

whether the model tends to under- or over-estimate the measured data, with an ideal value zero. However, positive and negative errors tend to cancel each other out. To prevent such cancellation, the Mean Square Error (Table 4, ID 4.4) criterion squares the residuals before calculating the mean, making all contributions positive and penalizing greater errors more heavily, perhaps reflecting the concerns of the user. The Root Mean Square Error (RMSE, Table 4, ID 4.5) takes the square root of the MSE to express the error metric in the same units as the original data. A similarly motivated measure is the Mean Absolute Error (MAE, Table 4, ID 4.6), but MSE and RMSE are usually preferred because they are smooth functions of the residuals, a requirement for many optimisation methods, whereas MAE has a kink at zero.

A complementary test, preserving the time dependence of the data is residual autocorrelation analysis. Assuming that any deterministic behaviour in the data is explained by the model, the remaining residuals should consist of white noise, i.e. with zero autocorrelation. So if the model "whitens" the residuals it can be reasonably assumed that all the deterministic behaviours have been included in the model. Conversely, a statistically non-zero autocorrelation for any lag >0 or a periodic behaviour (see Fig. 6) indicates non-white residuals induced by unmodelled behaviours. This reasoning can be summarized in Fig. 8.

4.2.3. Relative error and error transformations

In some studies, all events are not equally relevant for use as information to support decisions, designs, or interpretation; for example, in hydrologic modelling, one may be interested in either low-flow or high-flow conditions. Extremes may be of particular interest or of none, and may well dominate computed measures. Transforming the data or errors allows a focus on the aspects of interest.

Relative errors, *error/measured value*, weight the metric towards smaller values since larger ones may only have small relative error. The majority of metrics already defined can be calculated on relative errors, as in Table 5. Another option is to transform the residuals with a standard mathematical function to accentuate aspects of interest. For example, instead of squaring in

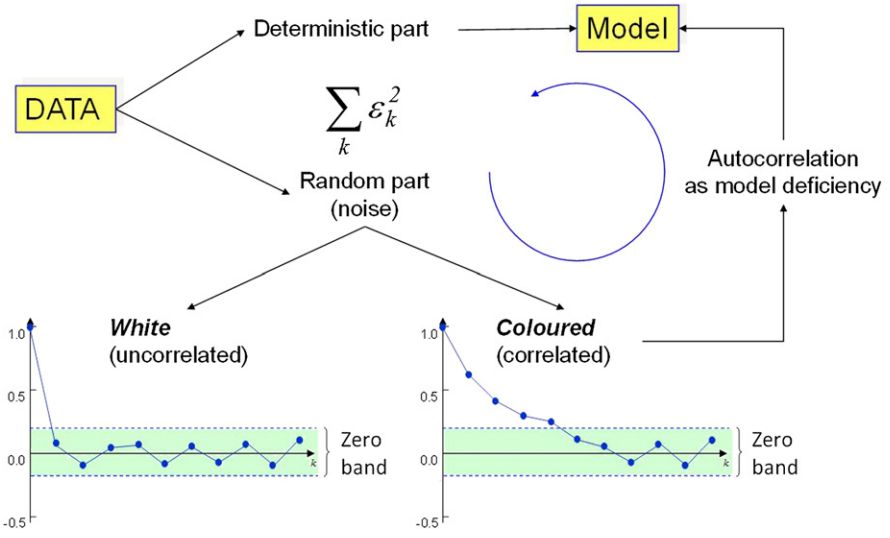

Fig. 8. Iterative process aimed at the whitening of the residuals. Autocorrelograms of white and coloured residuals are shown in the lower part. Autocorrelation is considered to be zero if its samples lie in the statistically zero band with limits $\pm\frac{1.96}{\sqrt{N}}$ where N is the number of experimental data points.

Table 5
Residual methods that use data transformations.

ID	Name	Formula	Notes
5.1	Relative bias	$\frac{1}{n}\sum_{i=1}^{n}\frac{((y_i+\varepsilon)-(\widehat{y}_i+\varepsilon))}{(y_i+\varepsilon)}$ $\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\widehat{y}_i}{y_i}$	Relative equivalent of ID 4.3, which increases the weighting of errors relating to low measurement values (e.g., low flow conditions in hydrological modelling). Ideal value is ≈ 0 and range $\pm\infty$. ε is a small value required in the event of $y_i = 0$.
5.2	Relative MSE (MSRE)	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(y_i+\varepsilon)-(\widehat{y}_i+\varepsilon)}{(y_i+\varepsilon)}\right)^2$	Relative equivalent of 4.4, Calculates the mean of the relative square errors. Ideal value is ≈ 0 and range $[0, \infty)$.
5.3	Fourth Root Mean Quadrupled (Fourth Power) Error (R4MS4E)	$\sqrt[4]{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^4}$	This approach modifies the RMSE by using the fourth power. This weights the error calculation towards large events within the record. Ideal value is 0 and range $(0, \infty)$.
5.4	Square-Root Transformed Root Mean Square Error (RTRMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\sqrt{y_i}-\sqrt{\widehat{y}_i})^2}$	RTRMSE uses the RMSE method (4.5), but in this case the data is pre-transformed by the square root function, to weight the error function towards lower values. The ideal value is still 0 and range $[0, \infty)$.
5.5	Log Transformed Root Mean Square Error (LTRMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i+\varepsilon)-\log(\widehat{y}_i+\varepsilon))^2}$	In a similar fashion the data in this case are pre-transformed by taking the logarithm of the data, which increases the weighting towards small values. It is important to note that this does not handle zero data well. The easiest way to overcome this is to add a very small ($\varepsilon = 1e-6$) value to each data point. Due to offset, ideal value is $= 0$ and range $(0, \infty)$.
5.6	Inverse Transformed Root Mean Square Error (ITRMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}((y_i+\varepsilon)^{-1}-(\widehat{y}_i+\varepsilon)^{-1})^2}$	The Inverse transform operates in a similar method to those previously mentioned. This method has the largest weighting towards small values and as in 5.5 it needs zero elements to be dealt with.
5.7	Relative MAE (MARE)	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{(y_i+\varepsilon)-(\widehat{y}_i+\varepsilon)}{(\widehat{y}_i+\varepsilon)}\right\rvert$	Relative equivalent of the mean absolute error (4.6) Due to offset, ideal value is $= 0$ and range $(0, \infty)$.
5.8	MdARE Median Absolute Percentage Error	$\text{median}\left(\left\lvert\frac{(y_i+\varepsilon)-(\widehat{y}_i+\varepsilon)}{(\widehat{y}_i+\varepsilon)}\right\rvert\right)\times 100$	This approach is similar to 5.7 but it uses the median to reduce the possible effect of outliers.
5.9	RVE Relative Volume Error	$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)}{\frac{1}{n}\sum_{i=1}^{n}(y_i)}$	The relative volume error compares the total error to the total measurement record. Similar to bias measurements, a low value does not mean low errors, just balanced errors.
5.10	Heteroscedastic Maximum Likelihood Estimator (HMLE)	$\frac{\frac{1}{n}\sum_{i=1}^{n}w_i(\lambda)[\widehat{y}_i-y_i]^2}{\left[\prod_{i=1}^{n}w_t(\lambda)\right]^{\frac{1}{n}}}$	Where the weights are typically calculated by $w_i = f_i^{2(\lambda-1)}$ with f_i the expected value of y_i (either y_i or $\widehat{y}_i$ can be used). λ is an *a priori* unknown shaping parameter necessary to stabilise the variance and is often adjusted along with the parameters (Sorooshian and Dracup, 1980).

RMSE, the fourth power could be used to accentuate large events, while the square root, logarithm and inverse functions could be used to accentuate small values. Formulae are presented for modifications to the RMSE (Table 5), applicable also to other criteria. For a discussion on the impact of transformations see Pushpalatha et al. (2012).

Another criterion weighting the residuals is maximum likelihood. Maximum-likelihood estimation (MLE) relies on assumptions about the distribution of the residuals, often that they are normally distributed, zero-mean and constant-variance. When assuming a heteroscedastic variance (proportional to the magnitude of events) this criterion becomes the HMLE proposed by Sorooshian and Dracup (1980). This reduces the influence of large events and provides more consistent performance over all event ranges. A formal likelihood function for correlated, heteroscedastic and non-Gaussian errors was recently proposed by Schoups and Vrugt (2010).

Relative (e.g., volume) error or relative bias of a predicted attribute compares the total error to the sum of the observations of the attribute over the time period or spatial extent of interest. Median absolute percentage error (Table 5, ID 5.8) modifies the mean absolute relative error (MARE) (Table 5, ID 5.7) by using the median instead of the mean.

4.3. Preserving the data pattern

Methods that consider each event in time or space as a separate item subsequently lose the evident patterns that often exist in time- and space-dependent environmental data. Adjacent values can tend to be strongly correlated, in which case this structure should be taken into account in the model. To test the ability of the model to preserve the pattern of data, performance metrics must include consideration of how data points and how their errors relate to each other.

A simple quantitative and graphical measure is the *cross-correlation* between measured and calculated values. It measures how the similarity of the two series varies with delay along one dimension (usually time). A simple standard deviation test is calculated to determine if behaviours are significantly similar. In addition, the cross-correlation can be computed between input data and residuals (Fig. 9); if a significant correlation is detected, then it will indicate that some behaviour of the system is not being accurately represented by the model. For some models (in particular statistical or regression-based), it is also appropriate to calculate the autocorrelation of the residuals. Significant autocorrelation may indicate unmodelled behaviour.

Perhaps the best known item in this category is the correlation coefficient. It is used to indicate how variation of one variable is explained by a second variable, but it is important to remember it does not indicate causal dependence. The Pearson Product-Moment Correlation Coefficient (PMCC), which calculates the correlation between two series of sampled data and lies between -1 and 1, is commonly used for model evaluation. Coefficient of Determination (r^2) is a squared version of PMCC that is also commonly used to measure the efficiency of a model, but only varies between 0 and 1 (Table 6).

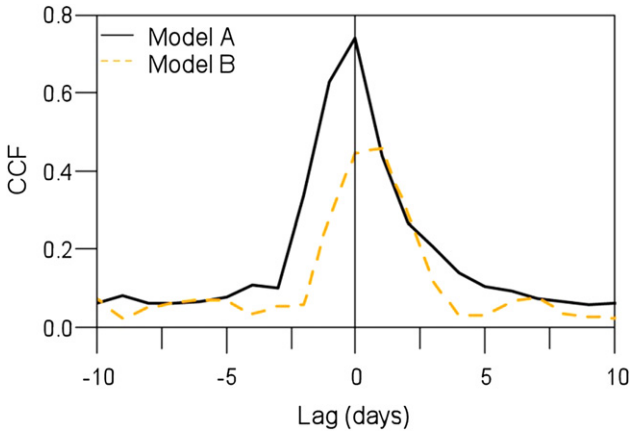

Fig. 9. Cross-correlation function (CCF) plot between effective rainfall and the model residuals for a hydrological model. Model A shows a significantly larger correlation between the residuals and input indicating there is more unmodelled behaviour in Model A. Data were generated from the IHACRES model.

In hydrologic modelling, the Coefficient of Determination is commonly known as the Nash–Sutcliffe efficiency coefficient (NSE or R^2; we use the former notation in this paper to avoid confusion with r^2) (Table 6, ID 6.1) (Nash and Sutcliffe, 1970). It indicates how well the model explains the variance in the observations, compared with using their mean as the prediction. A value of unity indicates a perfect model, while a value below zero indicates performance worse than simply using the mean. This criterion was shown to be a combination of other simple metrics (observed and simulated means and standard deviations (Table 2, ID 2.2 and 2.6) and the correlation coefficient (Table 6, ID 6.3)) by Gupta et al. (2009). The persistence index (e.g., Kitanidis and Bras, 1980) is similar to NSE, but instead of using the mean as the predictor variable, it uses the previous observed value as the predictor and is therefore well suited in a forecasting context.

Another criterion similar to NSE is the RMSE-standard deviation ratio (RSR), which weights the standard RMSE by the standard deviation of the observed values and is equivalent to the root square of 1 minus NSE. This standardises the criterion, allowing consistent evaluation of models when applied to systems with different variance inherent in the observed data (Moriasi et al., 2007).

All the evaluation methods mentioned above suffer potential bias. A model may have a significant offset and still yield ideal values of these metrics. The Index of Agreement (IoAD) compares the sum of squared error to the potential error. Potential error is the sum of squared absolute values of the differences between the predicted values and the mean observed value and between the observed values and the mean observed value. IoAD is similar to the coefficient of determination, but is designed to handle differences in modelled and observed means and variances.

Table 6
Correlation and model efficiency performance measures.

ID	Name	Formula	Range	Ideal value	Notes
6.1	Coefficient of determination/ Nash–Sutcliffe Model Efficiency (NSE)	$1-\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$	$(-\infty, 1)$	1	This method compares the performance of the model to a model that only uses the mean of the observed data. A value of 1 would indicate a perfect model, while a value of zero indicates performance no better than simply using the mean. A negative value indicates even worse performance.
6.2	Cross-Correlation Function (CCF)	ccf at lag n : $\sum_{i=1}^{\infty} u_i(y_{i+n}-\hat{y}_{i+n})$ $\sum_{i=1}^{\infty} y_i(y_{i+n}-\hat{y}_{i+n})$ acf at lag n : $\sum_{i=1}^{\infty}(y_i-\hat{y}_i)(y_{i+n}-\hat{y}_{i+n})$			Cross-correlation plots (Fig. 9) plot the cross correlation as a function of lag, including significance lines. Note that the correct formula for the cross correlation includes the complex conjugate of the first term, but this is not needed here as the quantities are real. In the first case look for any relationship between input (u) and residuals (there should be none), in the second case look for a shift between the modelled and measured CCF peaks (there should be none).
6.3	PPMC	$\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}}$	$(-1, 1)$	1	The Pearson Product moment correlation measures the correlation of the measured and modelled values. Negatives to this model are linear model assumptions and the fact it can return an ideal result for a model with constant offset.
6.4	RSqr (r^2) Coefficient of determination	$\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}}\right)^2$	$(0, 1)$	1	Squared version of 6.3, with the same interpretation of results, except range is now (0, 1).
6.5	IoAd Index of Agreement	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(\lvert\hat{y}_i-\bar{y}\rvert+\lvert y_i+\bar{y}\rvert)^2}$	$(0, 1)$	1	This method compares the sum of squared error to the potential error. This method is similar to 6.4 however it is designed to be better at handling differences in modelled and observed means and variances. Squared differences may add bias to large data value events (Willmott, 1981).
6.6	PI Persistence Index	$1-\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_{i-1})^2}$	–	–	The persistence index compares the sum of squared error to the error that would occur if the value was forecast as the previous observed value. Similar to 6.1 except the performance of the model is being compared to the previous value.
6.7	RAE Relative Absolute Error	$\frac{\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert}{\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\bar{y}\rvert}$	$(0, \infty)$	0	This compares the total error relative to what the total error would be if the mean was used for the model. A lower value indicates a better performance, while a score greater than one indicates the model is outperformed by using the mean as the prediction.
6.8	RSR RMSE – Standard deviation ratio	$\frac{\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$	$(0, \infty)$	0	The traditional RMSE method weighted by the standard deviation of the observed values (Moriasi et al., 2007).

When modelling a chaotic system, it may be impossible to reproduce an observed time-series, but still possible and desirable to reproduce observed data patterns. In this case, none of the error metrics presented in Table 6 will be appropriate. However, most of the direct value comparisons presented in Table 2 are still relevant. In addition, graphical comparisons of data patterns, such as auto-correlation plots and phase space plots may be useful in comparing the model with observations. In recent years, some authors have proposed a new family of visual performance measures that are based on attempts to mimic how the eye evaluates proximity between curves (Ewen, 2011; Ehret and Zehe, 2011). These criteria seem very promising, as they avoid traducing model errors simply in terms of difference of magnitude, and also include time shifts.

4.4. Indirect metrics based on parameter values

Model identification techniques (Ljung, 1999; Walter and Pronzato, 1997; Norton, 2009) deal with, among other things, how well the model parameters are identified in calibration. These techniques help to judge whether a model is over-fitted, i.e. has too many parameters relative to observed behaviour (e.g., Jakeman and Hornberger, 1993), and consequently how reliable is the computed value of the parameters. Attention to model identification is an important consideration in performance evaluation.

Some of these techniques, especially for models that are linear in the parameters, give the estimated variance of each parameter value, and more generally their joint covariance. Although estimated variance is a tempting indicator of the quality of a parameter estimate, it has to be treated with care. Low estimated parameter variance signifies only that estimates obtained in an ensemble of experiments using data with identical statistical properties would have small scatter. It is only as reliable as the estimates of those properties are. Moreover, variance alone says nothing about possible bias; mean-square error is variance *plus bias squared*. There is also a risk that, although all parameter variance estimates are small, high covariance exists between the estimates of two or more parameters, so their errors are far from independent. Hence some parameter combinations may be poorly estimated. The covariance matrix shows directly the correlations between all pairs of parameter estimates, but eigen-analysis of the covariance matrix is necessary to check for high estimated scatter of linear combinations of three or more parameter estimates. Keeping these limitations in mind, high estimated covariance may suggest that the model has too many parameters or that the parameters cannot be well identified from the available data. Uncertainty in the parameters, indicated by high covariance, may translate to high uncertainty in model-based predictions. In one example, instantaneous-unit-hydrograph parameters (in modelling of linear dynamics from effective rainfall to observed flow) are estimated by an automatic procedure from the data and have standard errors which can be used to estimate prediction variance (Young et al., 1980). This applies to any identification procedure which estimates parameter covariance. If this is not possible, then it is necessary to estimate the variance in other ways such as bootstrap or sensitivity analysis, which over a given parameter range will observe the changes in the objective function, from which the parameter variance can be estimated.

Other examples of parameter-based metrics are the Akaike Information Criterion (AIC) (Akaike, 1974), the similar Bayesian Information Criteria (Schwarz, 1978), originally designed for linear-in-parameters models to prevent over-fitting with too many parameters, and the Young Information Criterion (YIC; Young, 2011) (see Table 7). They allow determination of relative ranking between models, but cannot be used to evaluate how well the model approximates data in absolute terms. Weijs et al. (2010) endorse the use of information theory and corresponding scores for evaluating models, arguing that such approaches maximise the information extracted from observations.

The Dynamic Identifiability Analysis (DYNIA) approach developed by Wagener et al. (2003) is another parameter-based method. It uses a Monte Carlo approach to examine the relation between model parameters and multiple objective functions over time. Uncertainty analysis is performed over a moving time window to determine the posterior distributions of the parameters, allowing for the identification of 'informative regions' (Fig. 10) or potential structural errors in the model. Different optimal parameters at different times might indicate the model is failing to represent all modes of behaviour of the system.

Identifiability is a joint property of the model and the input/output data, concerning whether its parameters can be found with acceptable uncertainty (Norton, 2009, Section 8.2). Its analysis poses two questions. First, is the model identifiable in the sense that all the desired parameters would be found uniquely from noise-free data by error-free estimation? It would not be if, for instance, a transfer-function model could be identified but did not uniquely determine the physically meaningful parameters. That could be due to it having too few parameters or due to ambiguity, with no way to decide e.g., which of its poles related to which physical process. Bellman and Astrom (1970) introduced the idea as structural identifiability, although a better name is deterministic identifiability, as it depends on the data as well as the model structure. Walter (1982) gives an approach for comprehensive analysis of deterministic identifiability of models linear in their parameters. Analysis for non-linear models is much harder (Pohjanpalo, 1978; Holmberg, 1982; Chapman and Godfrey, 1996; Evans et al., 2002). Tests linking identifiability with noisy data to

Table 7
Some metrics based on model parameters.

ID	Name	Formula	Description
7.1	ARPE Average Relative Parameter Error	$\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\sigma}_i^2}{a_i}$	$\widehat{\sigma}_i^2$ is the estimated variance of parameter, a_i. Gives an indication of how well model parameters are identified and whether there is over-fitting. Range is (0, ∞) and ideal value is 0.
7.2	DYNIA Dynamic Identifiability Analysis	–	Localised sensitivity analysis to determine parameter identifiability by multiple objective functions. Determines optimal parameter sets and parameter identifiability (Wagener et al., 2003)
7.3	AIC Akaike Information Criterion	$m\log_n(\mathrm{RMSE}) + 2p$	This metric weights the RMSE error based on the number of points used in calibration, m, and the number of free parameters, p. The aim of this metric is to find the simplest model possible and prevent over-fitting (Akaike, 1974).
7.4	BIC Bayesian Information Criterion	$m\log_n(\mathrm{RMSE}) + p\log_n(m)$	A modification of AIC (7.3) (Schwarz, 1978)
7.5	YIC Young Information Criterion	$\mathrm{YIC} = \log_e\frac{\sigma^2}{\sigma_{\hat{y}}^2} + \log_e\mathrm{EVN}$ $\mathrm{EVN} = \frac{1}{n_\rho}\sum_{i=1}^{i=n_\rho}\frac{\widehat{\sigma}^2 p_{ii}}{\widehat{\rho}_i^2}$	Where n_ρ is the number of parameters in the ρ vector, p_{ii} is the ith diagonal element of the $\mathbf{P}(N)$ matrix obtained from the estimation analysis. YIC combines the residual variance of the model and parameter efficiency (Young, 2011).

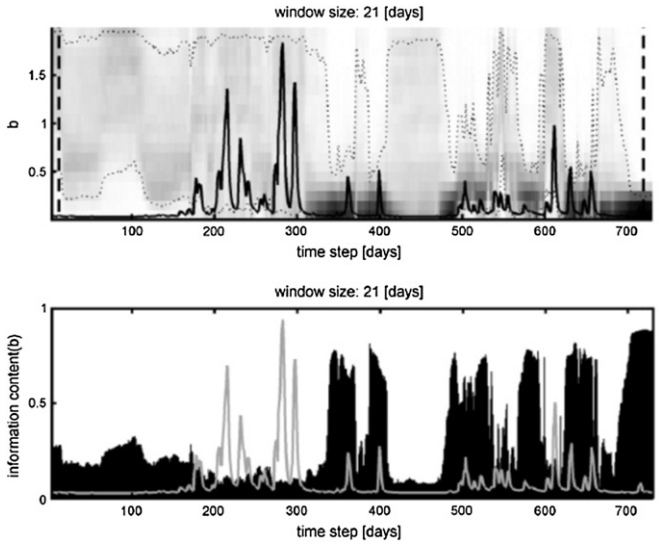

Fig. 10. Example DYNIA plots. The top plot shows for a parameter (b), the temporal change of the marginal posterior distribution against the observed stream flow. The bottom plot shows the temporal variation of data information content with respect to one of the model parameters. Taken from Wagener and Kollat (2007).

sensitivity in Monod-like kinetics were presented by Vanrolleghem and Keesman (1996), Dochain and Vanrolleghem (2001), Petersen et al. (2003) and Gernaey et al. (2004).

The second question in testing identifiability is whether the uncertainties in the data allow adequate estimation of the parameters. The properties of the input are crucial here, in particular whether persistency-of-excitation conditions are met (Ljung, 1999; Norton, 2009, Section 8.3), ensuring that the system behaviour represented by the model is fully excited. A further critical aspect is the sampling scheme for the input and output, as a too-low sampling rate may prevent identification of rapid dynamics or, through aliasing, lead to them being misidentified, while a too-high sampling rate may result in avoidable numerical ill-conditioning in the estimator. Finally, the presence of feedback may determine identifiability (Norton, 2009, Section 8.4).

The relationship between identifiability and sensitivity through the Fisher Information Matrix (FIM) was explored in depth by Petersen (2000) and De Pauw (2005), and later applied to river-quality models by Marsili-Libelli and Giusti (2008).

Confidence Region Analysis is a very effective tool for assessing the reliability of parametric identification. It is based on the computation of the covariance matrix of the estimated parameters (Bates and Watts, 1988; Seber and Wild, 1989; Marsili-Libelli, 1992; Dochain and Vanrolleghem, 2001) by differing means and on inferring the model consistency by their agreement or divergence (Marsili-Libelli et al., 2003; Checchi and Marsili-Libelli, 2005; Marsili-Libelli and Checchi, 2005). The divergence between the estimated confidence regions computed in the exact way (Hessian matrix) and through a linear approximation (FIM matrix) indicates poor parameter identification.

4.5. Data transformation methods

The transformation of data into different domains is another technique that can be used for performance evaluation. Transformation can often highlight aspects of a model's behaviour that were not clear in the original time or space domain.

4.5.1. Fourier transformation

The most common and best-known transformation is the Fourier transform which converts data into the frequency domain (Table 8, ID 8.1). The Fourier transform represents the original domain vector as a series of complex exponentials, where each complex exponential can be interpreted as representing a frequency of the signal (variable) with a magnitude and phase. Results can be plotted (Fig. 11) to allow modelled and observed data to be directly compared in the frequency domain. Here, transforming the data can give insights into model performance that might not be obvious in untransformed data – for instance, the model may be accurately representing seasonal transitions but missing some effect that occurs on monthly time-scales. Another option is to sample model output and observed data at specific frequencies and calculate their differences. The Fourier transform is commonly evaluated in terms of the power spectrum, which squares the magnitude of the transform; the power spectrum is the Fourier transform of the autocorrelation function (Ebisuzaki, 1997).

4.5.2. Wavelets

An extension of the Fourier transformation uses wavelets. Unlike the Fourier transform, which uses complex exponentials defined over the entire original domain (time or space), wavelet transforms use functions with a finite energy, allowing the transformation to be localised in frequency and temporal or spatial location (Table 8, ID 8.2 and 8.3) adapting the resolution of the analysis, both in time and scale, to each portion of the signal. As a result, model performance can be quantitatively assessed on different temporal and spatial scales, using wavelets to separate

Table 8
Details of transformation methods.

ID	Name	Formula	Description
8.1	Fourier transform and power spectral density	$Y_k = \frac{1}{n}\sum_{i=1}^{n} y_i e^{-\frac{2\pi j}{n}ki}$ $I(\omega_k) = \frac{1}{n}\lvert Y_k\rvert^2$	Where Y is the Fourier transform, the same length as sequence y and j is the imaginary unit. Many algorithms are available to calculate the Fourier transform. Signals are commonly compared based on their power spectral density function, which can be estimated as $I(\omega_k)$ (Fig. 11).
8.2	Continuous wavelet transform	$C_i^y(s) = \sum_{i'=0}^{n-1} y_{i'} \Psi^*\left(\frac{i'-i}{s}\right)\delta t$	Where $\Psi$is the mother wavelet function, δt is the time step, i and s are the scale and place parameters. Many possible calculations can be made once the transformation is complete, please see Lane (2007) for complete details.
8.3	Discrete wavelet transform	$Y[n] = (y_i{*}g)[n] = \sum_{k=-\infty}^{k=\infty} y_i g_{n-i}$	Where n is the resolution level, g is the filter coefficients from the discrete wavelet (note that there will be two sets, one for the high pass and one for low pass filter). Each resolution level can then have performance criteria applied separately (Chou, 2007).
8.4	EOF Empirical Orthogonal Functions	$A(t,s) = \sum_{k=1}^{M} C_k(t) u_k(s)$	EOF analysis produces a result which specifies the space–time field as a linear combination of a spatial basis function, u_k, with expansion functions of time c_k, over multiple modes (M). Typically analysis will be completed for only the first couple of modes which explain the most variance. Need to compare both u_k and c_k to evaluate model performance. See Hannachi et al. (2007) for more details.

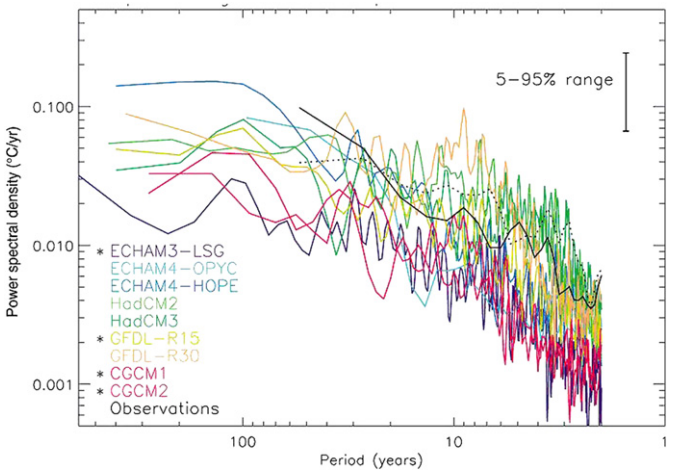

Fig. 11. Power spectral density graph of climate model global mean temperatures. The graph compares the frequency response of different climate models against observations, taken from IPCC (2001).

signals and remove interference across scales or between locations. Wavelet transforms are apt for assessing the temporal and spatial variability of high-frequency data, such as *in situ* automated sensor data or remote sensing data (Kara et al., 2012).

Wavelet analysis (Strang and Nguyen, 1996; Torrence and Compo, 1998) is based on adapting a set of replicas of an initial function (mother wavelet) of finite duration and energy by adapting it to the data through shifting (change of time origin) and scaling (change of amplitude). The use of wavelet-based transforms to analyse model performance will be highlighted using two contrasting examples. In the first, Lane (2007) uses a discrete approximation of the Continuous Wavelet Transform (CWT) to analyse performance of rainfall–runoff model results. In particular, the localisation feature allows the models to be analysed for errors that appear at certain times and locations. Once the CWT has been calculated, several types of analysis can be performed. For example, the spectrum of the CWT can be calculated (similar to the Fourier transform PSD), from which error metrics or cross-correlations can be calculated between the modelled and observed data (Fig. 12). Errors and potential time lags can be found from the power and phase data.

Continuous Wavelet Analysis involves analysing a huge number of coefficients and this redundancy is often confusing. By contrast, if the wavelet replicas by shifting and scaling are constrained to powers of 2 (dyadic decomposition), then the original time-series can be neatly split into low-frequency components (approximations) and high-frequency details, still retaining the double time-scale representation. Chou (2007) implements the stationary wavelet transform, a variant of the discrete wavelet transform (DWT), to perform multi-resolution analysis. At each step, the signal (the discrete wavelet in this case) is passed through specially designed high-pass and low-pass filters. Through a cascade of filters, the signal can be decomposed into multiple resolution levels. Chou (2007) decomposed the calculated and observed values into different time-scales; model performance could be quantitatively evaluated at each level. Examples of DWT for signal smoothing and denoising can be found in Marsili-Libelli and Arrigucci (2004) and Marsili-Libelli (2006).

4.5.3. Empirical orthogonal functions

Another approach to data transformation is to use Empirical Orthogonal Functions (EOF), which has been extensively used, for instance, for analysis of spatial models in meteorology, climatology and atmospheric science (see Hannachi et al., 2007 for a thorough review). EOF analysis extends principal-component analysis. For a space–time field, EOF analysis finds a set of orthogonal spatial patterns and an associated uncorrelated time series (or principal components). This allows the model to be judged on a spatial as well as temporal basis (Doney et al., 2007). The orthogonal spatial patterns explain the variance in the spatial model, allowing identification of the pattern that explains the most variance, which can be compared between modelled and measured values. It is also important to observe how well the model calculates the temporal (principal) components of the data. Methods have been developed for various model types and studies, including variations focussing solely on spatial resolution or patterns correlated in time.

5. Qualitative model evaluation

Despite the power of quantitative comparisons, model acceptance and adoption depend in the end strongly on qualitative, and often subjective, considerations. There may be even more subjectivity in considering the results of quantitative testing. Suppose one model is superior to another according to one metric, while it is the reverse according to another metric. What is the weight to be assigned to individual quantitative criteria in the overall assessment? How are these weights decided and by whom? Qualitative assessments may become essential in highly complex or data-poor situations, when data are scarce or unreliable. In such cases, we may be more interested in qualitative model behaviour that can sketch out trends and system behaviour, than in quantitative output that produces actual values for variables. These qualitative assessments typically involve experts, defined as those with appropriate extensive or in-depth experience in the domain, including non-professionals (Krueger et al., 2012).

Expert opinion can be elicited to provide endpoint estimates to compare against model results (Rowan et al., 2012; Giordano and Liersch, 2012), or estimates of uncertainty associated with model components (Page et al., 2012). A common form of expert evaluation involves peer review of the conceptual model to assess whether the logic and details of the model matches its intended

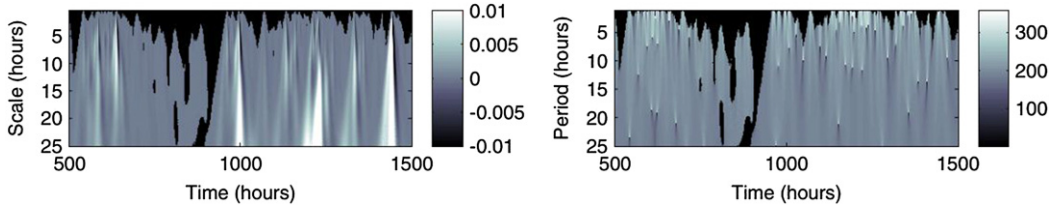

Fig. 12. Example of error plots created using the CWT contour plots of power error (the difference between wavelet power spectra with simulated error and the original data series) (left) and phase difference for simulations (right). The plots show model performance varying through the dataset and thus an appropriate performance indicator is one that captures this internal variability. Taken from Lane (2007).

purpose (Nguyen et al., 2007; Krueger et al., 2012). Expert opinion is inherently uncertain and experts even from the same domain may hold diverse, possibly opposing, viewpoints. Methods are available for aggregating or finding consensus among multiple opinions, for example, through evaluating the level of mutual agreement between opinions (Vrana et al., 2012). However there are cases where consensus is not the objective, and rather, important information lies within the disagreement among experts (Knol et al., 2010).

In some cases a model is valued not for the accuracy of its predictive power, but by other outcomes, including community- and capacity-building, and the ontological and educational functionality that it brings to groups of stakeholders or users who gain from being part of modelling process (Voinov and Bousquet, 2010; Krueger et al., 2012). There have been various attempts to evaluate the modelling process and its usefulness for stakeholders (Campo et al., 2010). Questions of construction, operation and output of the model that may be relevant to its adoption cover a number of issues (Robson et al., 2008; Parker et al., 2002; Risbey et al., 1996), ranging from uncertainty in model components to unexpected behaviour and to how it is presented to users. Relevant questions are:

- Has the user community been clearly identified?
- Does the model meet its specified purpose?
- How reliable are the input data (e.g., measurement error, sampling rate)? How might this affect model output?
- Does the model behave as expected? How is the model behaviour affected by assumptions required in development of the model?
- Is the model structure plausible? Have alternative model structures/types been tested?
- Has the model been calibrated satisfactorily?
- Is the model flexible/transparent enough for the intended users?
- Does the model improve users' ability to understand system behaviour (compared with no model or with a simpler model)?
- Is the model useful as a co-learning tool and can it help reconcile or synchronize knowledge between various users or stakeholders?
- Does the model tell us anything new about the system? Are there any emergent properties that deserve further attention or that explain known phenomena?

Many of these questions can be answered only qualitatively, but their answers may be more important in practice than quantitative performance evaluation. In some modelling communities, this has led to the development of systematic protocols to ensure 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.

As demonstrated by Seppelt and Richter (2005, 2006), it is possible for a particular numerical procedure to produce different solutions with different modelling packages due to internal coding. This highlights the importance of careful treatment of model results and qualitative forms of evaluation. Sometimes it is useful to assign numerical values to the qualitative answer of each question, allowing the models to be evaluated numerically or graphically. One system using this approach is the Numerical Unit Spread Assessment Pedigree (NUSAP) system (van der Sluijs et al., 2004). This system combines derived numerical metrics (including some form of error calculation and spread calculation) with more qualitative approaches to assess the performance of the model and the process generating the model. The results from these multi-criteria tests are combined onto a single kite-diagram allowing easy comparison of performance for multiple models (Fig. 13). A similar methodology was adopted by Devisscher et al. (2006) to assess the model performance of differing wastewater treatment control schemes.

6. Performance evaluation in practice: a suggested general procedure

As pointed out by many authors (e.g., Jakeman et al., 2006), performance evaluation is just one step of iterative model development. Evaluation results may indicate whether additional study is necessary. If performance is unsatisfactory, then different data, calibration procedures and/or model structures should be considered. With satisfactory performance, one may also evaluate whether simplification or other modification would entail significant performance loss. Modelling is an iterative process and model evaluation is as well. Model evaluation occurs repeatedly at various stages, helping the model developers to keep the modelling process 'in check' and 'under control'.

A question that must be addressed when using any of the performance metrics discussed in Section 4 is: what constitutes a "good" or "acceptable" match? During parameter estimation, the aim may be simply to select the model with the lowest error, whatever its value, but for final evaluation it is often desirable to have some more objective criterion. This may be a pre-determined value that an end-user has determined will allow confident decision-making, or it may be derived by comparison with the accuracy of other models in a similar setting, such as the statistics presented by Arhonditsis and Brett (2004) for aquatic

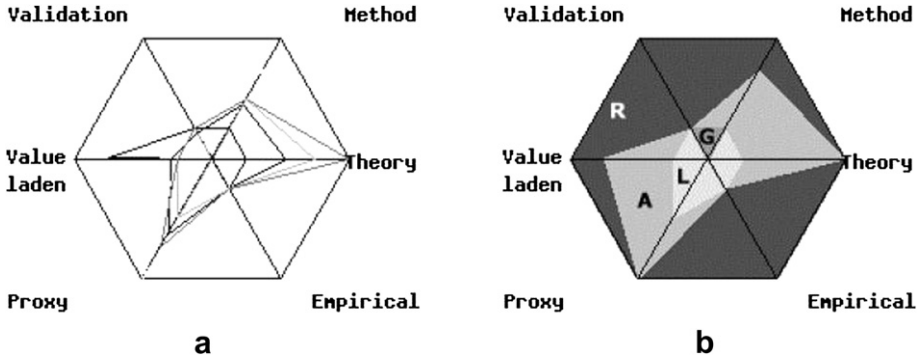

Fig. 13. Results from the NUSAP system. Numerical values are assigned to qualitative fields and a simple kite diagram is used for multi-criteria evaluation (van der Sluijs et al., 2004, 2005).

biogeochemical models. A few authors have attempted to relate expert judgement and numerical criteria (e.g., Chiew and McMahon, 1993; Houghton-Carr, 1999; Crochemore, 2011).

Every modelling endeavour has unique goals and challenges, so there is no ideal or standard technique for evaluation that can be applied for all models. Even so, it is possible to suggest some general elements that are beneficial in modelling processes as summarised in the five steps below. This procedure assumes that the model is already built. Also, the steps can be linked to existing modelling guidelines (e.g., Henriksen et al., 2009; Hutchins et al., 2006; Piuleac et al., 2010; Blocken and Gualtieri, 2012).

Step 1. Reassess the model's aim, scale, resolution and scope

The most important step is to have a clear idea of the modelling purpose: what the model is used for, the decision to be taken, the understanding we want to improve. This means in turn having a clear idea of what conditions are being modelled and determining what constitutes a 'good' model. In a hydrologic modelling example, is the modelling objective to reproduce high-magnitude events or the entire range of events? Is it more important for the model to replicate aggregated (e.g., monthly, seasonal, annual) behaviour or individual, perhaps critical, events? Having a clear idea of the model's purpose allows a first selection of error metrics. It helps ensure the metrics are suited to the context of the model construction, including its assumptions, scales and resolution. Often the model's aim and scope changes during the model life cycle and evaluation of performance metrics can sometimes guide the study in other directions. In order to keep track of model changes, reflect on its context and communicate its purpose to new users, it may help to document the model description in a standardised way, such as that described by Grimm et al. (2006, 2010).

Step 2. Check the data

The second step determines whether sufficient data are available and how they can be split between calibration and performance evaluation. For initial data analysis, a graphical procedure is suggested to examine the general behaviour of the data. For time series data, autocorrelation will detect any periodicity or drift, while calculating the empirical distribution function will give a better impression of the magnitude of events. It is desirable to examine a time–domain plot of events to detect periods in which events and outliers occur. Autocorrelation can also be applied to spatial datasets to identify spatial periodicity (e.g., turbulent eddies) or inhomogeneity. After these tests, the data pattern and subsets for calibration and testing can be selected with higher confidence. Subsequent model evaluations must balance the limits of computing resources and time. For example, a modeller may decide to reshuffle the data and increase the calibration domain, often times at the expense of the quality of further performance evaluations that will be possible.

Step 3. Visual performance analysis

The third step entails visual analysis to judge the model performance, using graphics, maps, animations and other digital methods for viewing patterns in data (see Kelleher and Wagener, 2011). There are two main goals: the detection of under- or non-modelled behaviour and gaining an overview of the overall performance.

Unmodelled behaviour can be detected by means such as the residual plot, QQ plot and cross-correlation between the input data and residuals, all capable of indicating when a model is not representing a system's behaviour adequately. These results can be used to help refine the model before further evaluation.

In some contexts, simple graphical representation of model output is sufficient. The great strength of visualisation is that details can be observed in the results which would have remained hidden in a quantitative evaluation, or which can help to direct the tools used for quantitative evaluation. Visualisation takes advantage of the strong human capacity for pattern detection and may allow model acceptance or rejection without determining strict formal criteria in advance. Kuhnert et al. (2006), for instance, explore the output of a visual comparison in a web survey of 100 people, together with a variety of algorithms of spatial comparison.

Step 4. Select basic performance criteria

RMSE or r^2 are good initial candidates for a metric as their wide usage aids communication of the model performance. Thorough understanding of any weakness of the metric for the particular purpose is essential.

Even in initial model evaluation, multiple metrics should be considered in view of the weaknesses of individual metrics (e.g., Carpentieri et al., 2012; Smiatek et al., 2012). For example, the coefficient of determination r^2, which can suffer significant offset error, should be paired with bias. RMSE (or again r^2) can be paired with a selected data transformation to reduce the bias that some conditions (e.g., large events) introduce into the evaluation. A more elegant approach would be to use the KGE criterion proposed by Gupta et al. (2009), in which the bias is an explicit component.

Step 5. Refinements (and back to model improvements and eventually Step 1 at the next stage)

Once analysis has been completed, it is possible to consider how exhaustive the evaluation has been. The first set of metrics adopted is judged against the knowledge gained from visual analysis in Step 3, as well as how well the current evaluation differentiates between competing models or interpretations. Depending on the problems identified, for example, if changes in model divergence over time/space are not captured by the metrics, then a windowed metric or a more advanced wavelet analysis may be needed. When a significant difference between calibration and testing model performance is detected, calibration data/procedures may have been inadequate, and sensitivity analysis may help determine which parameters are causing problems. If significant divergences, for example between low/high-magnitude events, are not captured by metrics, then data transformations or multi-resolution methods to highlight the differences may be adopted. As already pointed out, these refinements may entail revision of the model structure and/or data, so the procedure may require additional cycle(s).

It is crucial to engage the user community in all these five steps to the extent possible. Even without the users understanding the methods in detail, dialogue maintained with users may help the modeller identify the most suitable performance metrics (Kämäri et al., 2006). In the end, the user or stakeholder determines whether model performance is acceptable and the choice of metric should capture the users' expectations.

As shown in the previous sections, methods developed in different application areas can inform practice across disciplines. More broadly, methodological knowledge from environmental modelling may benefit from a common classification to reconcile quantitative approaches to performance evaluation. Additionally, modern computer software tools allow incremental construction of a common repository of implemented methods to support performance evaluation of a large class of environmental models. At present, discipline specific examples are emerging (e.g., Olesen, 2005; Gilliam et al., 2005; Dawson et al., 2007, 2010), but a generalised repository of evaluation approaches is needed across the spectrum of environmental modelling communities. Such a repository should allow users to select only the methods that are of interest for their case (avoiding overload from redundant

information), should allow references to case studies on the application of different methods, and should be open to additions and discussion. Given the subjective and qualitative aspects of model evaluation alluded to above, it may constitute a real "community" tool for environmental modellers. It would be a major step in improving the characterisation of model performance so that we as a community can enunciate more explicitly the strengths and limitations of our models.

7. Conclusions

This paper provides an overview of qualitative and quantitative methods of characterising performance of environmental models. Qualitative issues are crucial to address. Not everything can be measured. However, quantitative methods that make use of data were emphasised as they are somewhat more objective and can be generalised across model applications. This includes not just methods that couple real and modelled values point by point, but also methods that involve direct value comparison, preserving data patterns, indirect metrics based on parameter values, and data transformations. We also stress that observed data are the result of selection and interpretation. Data for modelling should be assessed critically in terms of its information content, reliability and generality.

We proposed a structured workflow to characterise the performance of environmental models. The modeller critically assesses the model scope and purpose, and the quality of the data. A preliminary performance analysis with visual and other techniques provides an overview that leads to the selection of simple, general performance criteria. Appraisal of intermediate results may identify the need to revisit the previous steps and use more advanced methods, depending on model aims and scope.

The overview of methods and the suggested structured workflow address the growing demand for more standardisation in evaluation of environmental models (Alexandrov et al., 2011; Bellochi et al., 2009). This paper aims to provide a common base of methods with a supporting workflow, rather than a more detailed evaluation standard. It would be difficult to come up with such standards given that models are built for various purposes, and the goals of the modelling effort very much determine the characterisation of a model. What is good for one application may turn out to be inadequate for another. For example, a model that is performing quite poorly in terms of simulating daily dynamics can do a very good job in describing the annual trends, which could be exactly what is needed for a particular decision analysis. In principle an evaluation standard should then contain both a characterisation of a model and a description of the case study, the goal of the model.

No matter how the model is used, it is still always good to know how it performs compared to the datasets that are available. Our main message is that characterising model performance should be considered an iterative process of craftsmanship. The selection of tools for a given context requires both expertise and creative skill. To produce a model that fulfils both modeller and stakeholder expectations, cooperative dialogue is needed, perhaps crossing multiple disciplines or areas of knowledge. In order to guide this process, a modeller cannot afford to restrict themselves to one standard recipe. "If all you have is a hammer, everything looks like a nail". A modeller needs to have a firm grasp of the broad range of available methods for characterising model performance. We hope that this paper will be a useful contribution and reference.

References

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

Alexandrov, G.A., Ames, D., Bellocchi, G., Bruen, M., Crout, N., Erechtchoukova, M., Hildebrandt, A., Hoffman, F., Jackisch, C., Khaiter, P., Mannina, G., Mathunaga, T., Purucker, S.T., Rivington, M., Samaniego, L., 2011. Technical assessment and evaluation of environmental models and software: letter to the Editor. Environmental Modelling and Software 26 (3), 328–336.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.H., Valéry, A., 2009. Crash tests for a standardized evaluation of hydrological models. Hydrology and Earth System Sciences 13, 1757–1764. http://dx.doi.org/10.5194/hess-13-1757-2009.

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Marine Ecology-Progress Series 271, 13–26.

Bates, D., Watts, D., 1988. Nonlinear Regression Analysis and its Applications. John Wiley & Sons, New York.

Beck, B., 2006. Model Evaluation and Performance. In: Encyclopedia of Environmetrics. John Wiley & Sons, Ltd.

Bellman, R., Astrom, K.J., 1970. On structural identifiability. Mathematical Biosciences 7, 329–339.

Bellochi, G., Rivington, M., Donatelli, M., Matthews, K., 2009. Validation of biophysical models: issues and methodologies. A Review. Agronomy for Sustainable Development. http://dx.doi.org/10.1051/agro/2009001.

Bennett, N., Croke, B.F.W., Jakeman, A.J., Newham, L.T.H., Norton, J.P., 2010. Performance evaluation of environmental models. In: Swayne, D.A., Yang, W., Voinov, A.A., Rizzoli, A., Filatova, T. (Eds.), Proceedings of 2010 International Congress on Environmental Modelling and Software. iEMSs, Ottawa, Canada.

Berthet, L., Andréassian, V., Perrin, C., Loumagne, C., 2010. How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. Hydrological Sciences Journal 55 (6), 1063–1073.

Beven, K.J., 2008. Environmental Modelling: an Uncertain Future? An Introduction to Techniques for Uncertainty Estimation in Environmental Prediction. Taylor and Francis.

Blocken, B., Gualtieri, C., 2012. Ten iterative steps for model development and evaluation applied to Computational Fluid Dynamics for Environmental Fluid Mechanics. Environmental Modelling and Software 33, 1–22.

Boucher, M.A., Perreault, L., Anctil, F., 2009. Tools for the assessment of hydrological ensemble forecasts obtained by neural networks. Journal of Hydroinformatics 11 (3–4), 297–307.

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

Campo, P.C., Bousquet, F., Villanueva, T.R., 2010. Modelling with stakeholders within a development project. Environmental Modelling and Software 25 (11), 1302–1321.

Carpentieri, M., Salizzoni, P., Robins, A., Soulhac, L., 2012. Evaluation of a neighbourhood scale, street network dispersion model through comparison with wind tunnel data. Environmental Modelling and Software 37, 110–124.

Chapman, M.J., Godfrey, K.R., 1996. Nonlinear compartmental model indistinguishability. Automatica 32 (3), 419–422.

Checchi, N., Marsili-Libelli, S., 2005. Reliability of parameter estimation in respirometric models. Water Research 39, 3686–3696.

Chiew, F.H.S., McMahon, T.A., 1993. Assessing the adequacy of catchment streamflow yield estimates. Australian Journal of Soil Research 31, 665–680.

Choi, H., Beven, K., 2007. 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.

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

Christoffersen, P., 1998. Evaluating intervals forecasts. International Economic Review 39, 841–862.

Cochran, W.G., 1977. Sampling Techniques, third ed. Wiley.

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

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., Hendrickx, F., 2012. Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. Water Resources Research 48 (5), W05552. http://dx.doi.org/10.1029/2011WR011721.

Costanza, R., 1989. Model goodness of fit: a multiple resolution procedure. Ecological Modelling 47, 199–215.

Crochemore, L., 2011. Evaluation of Hydrological Models: expert Judgement vs Numerical Criteria. Master thesis, Université Pierre et Marie Curie, PolyTech Paris, France, 54 pp.

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

Dawson, C., Abrahart, R., See, L., 2010. HydroTest: further development of a web resource for the standardised assessment of hydrological models. Environmental Modelling and Software 25 (11), 1481–1482.

De Pauw, D.J.W., 2005. Optimal Experimental Design for Calibration of Bioprocess Models: a Validated Software Toolbox. PhD thesis in Applied Biological Sciences, BIOMATH, University of Gent. URL: http://biomath.ugent.be/publications/download/.

Devisscher, M., Ciacci, G., Fé, L., Benedetti, L., Bixio, D., Thoeye, C., De Gueldre, G., Marsili-Libelli, S., Vanrolleghem, P.A., 2006. Estimating costs and benefits of advanced control for wastewater treatment plants – the MAgIC methodology. Water Science and Technology 53 (4–5), 215–223.

Dochain, D., Vanrolleghem, P.A., 2001. Dynamical Modelling and Estimation in Wastewater Treatment Processes. IWA Publishing, London.
Doney, S.C., Yeager, S.G., Danabasoglu, G., Large, W.G., McWilliams, J.C., 2007. Mechanisms governing interannual variability of upper ocean temperature in a global ocean hindcast simulation. Journal of Physical Oceanography 37 (7), 1918–1938.
Ebisuzaki, W., 1997. A method to estimate the statistical significance of a correlation when the data are serially correlated. Journal of Climate 10, 2147–2153.
Ehret, U., Zehe, E., 2011. Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrology and Earth System Sciences 15 (3), 877–896.
Evans, N.D., Chapman, M.J., Chappell, M.J., Godfrey, K.R., 2002. Identifiability of uncontrolled nonlinear rational systems. Automatica 38 (10), 1799–1805.
Ewen, J., 2011. Hydrograph matching method for measuring model performance. Journal of Hydrology 408 (1–2), 178–187.
FAIRMODE, 2010. Guidance on the Use of Models for the European Air Quality Directive. Working Document of the Forum for Air Quality Modelling in Europe FAIRMODE. Version 6.2.
Fox, D.G., 1981. Judging air quality model performance. Bulletin of the American Meteorological Society 62 (5), 599–609.
Frey, C.H., Patil, S.R., 2002. Identification and review of sensitivity analysis methods. Risk Analysis 22 (3), 553–578. http://dx.doi.org/10.1111/0272-4332.00039.
Gernaey, K.V., van Loosdrecht, M.C.M., Henze, M., Lind, M., Jørgensen, S.B., 2004. Activated sludge wastewater treatment plant modelling and simulation: state of the art. Environmental Modelling and Software 19, 763–783.
Ghelli, A., Ebert, E., March 2008. Meteorological applications. Special Issue: Forecast Verification 15 (1).
Gilliam, R.C., Appel, W., Phillips, S., 2005. The atmospheric model evaluation tool: meteorology module. In: Proc. 4th Annual CMAS Models-3 Users' Conference, September 26–28, 2005, Chapel Hill, NC.
Giordano, R., Liersch, S., 2012. A fuzzy GIS-based system to integrate local and technical knowledge in soil salinity monitoring. Environmental Modelling and Software 36, 49–63.
Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.
Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Müller, M., Pe'er, G., Pioub, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R.A., Vabø, R., Visser, U., DeAngelis, D.L., 2006. A standard protocol for describing individual-based and agent-based models. Ecological Modelling 198 (1–2), 115–126. http://dx.doi.org/10.1016/j.ecolmodel.2006.04.023.
Grimm, V., Berger, U., DeAngelis, D.L., Polhill, J.G., Giske, J., Railsback, S.F., 2010. The ODD protocol: a review and first update. Ecological Modelling 221 (23), 2760–2768. http://dx.doi.org/10.1016/j.ecolmodel.2010.08.019.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. Journal of Hydrology 377 (1–2), 80–91.
Gupta, H.V., Clark, M.P., Vrugt, V.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resources Research W08301. http://dx.doi.org/10.1029/2011WR011044.
Gy, P., 1998. Sampling for Analytical Purposes. Wiley, New York.
Haefner, J.W., 2005. Modeling Biological Systems, Principles and Applications. Springer, New York, 475 pp.
Hannachi, A., Jolliffe, I.T., Stephenson, D.B., 2007. Empirical orthogonal functions and related techniques in atmospheric science: a review. International Journal of Climatology 27 (9), 1119–1152.
Hejazi, M.I., Moglen, G.E., 2008. The effect of climate and land use change on flow duration in the Maryland Piedmont region. Hydrological Processes 22 (24), 4710–4722.
Henriksen, H., Refsgaard, J., Højberg, A., Ferrand, N., Gijsbers, P., Scholten, H., 2009. Harmonised principles for public participation in quality assurance of integrated water resources modelling. Water Resources Management 23 (12), 2539–2554.
Holmberg, A., 1982. On the practical identifiability of microbial models incorporating Michaelis–Menten-type nonlinearities. Mathematical Biosciences 62, 23–43.
Houghton-Carr, H.A., 1999. Assessment criteria for simple conceptual daily rainfall–runoff models. Hydrological Sciences Journal 44 (2), 237–261.
Hutchins, M., Urama, K., Penning, E., Icke, J., Dilks, C., Bakken, T., Perrin, C., Saloranta, T., Candela, L., Kämäri, J., 2006. The model evaluation tool: guidance for applying benchmark criteria for models to be used in river basin management. Large Rivers 17 (1–2), 23–48.
IPCC, 2001. In: Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Xiaosu, D. (Eds.), Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change (IPCC). Cambridge University Press, UK, p. 944.
Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall–runoff model? Water Resources Research 29 (8), 2637–2649.
Jakeman, A.J., Littlewood, I.G., Whitehead, P.G., 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. Journal of Hydrology 117, 275–300.
Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software 21 (5), 602–614.
Kämäri, J., Boorman, D., Icke, J., Perrin, C., Candela, L., Elorza, F.J., Ferrier, R.C., Bakken, T.H., Hutchins, M., 2006. Process for benchmarking models: dialogue between water managers and modellers. Large Rivers 17 (1–2), 3–21.
Kara, E.L., Hanson, P., Hamilton, D., Hipsey, M.R., McMahon, K.D., Read, J.S., Winslow, L., Dedrick, J., Rose, K., Carey, C.C., Bertilsson, S., da Motta Marques, D., Beversdorf, L., Miller, T., Wu, C., Hsieh, Y.-F., Gaiser, E., 2012. Time-scale dependence in numerical simulations: assessment of physical, chemical, and biological predictions in a stratified lake at temporal scales of hours to months. Environmental Modelling and Software 35, 104–121.
Keesman, K.J., Koskela, J., Guillaume, J.H., Norton, J.P., Croke, B., Jakeman, A., 2011. Uncertainty modelling and analysis of environmental systems: a river sediment yield example. In: Proceedings of MODSIM2011 International Congress on Modelling, Perth, Australia, 2011.
Kelleher, C., Wagener, T., 2011. Ten guidelines for effective data visualization in scientific publications. Environmental Modelling and Software 26 (6), 822–827.
Kitanidis, P.K., Bras, R.L., 1980. Real-time forecasting with a conceptual hydrologic model. 2. Application and results. Water Resources Research 16 (6), 1034–1044.
Kleijnen, J.P.C., Bettonvil, B., Van Groenendaal, W., 1998. Validation of trace-driven simulation models: a novel regression test. Management Science 44 (6), 812–819.
Kleijnen, J.P.C., 1999. Validation of models: statistical techniques and data availability. In: Proceedings of the 31st Conference on Winter Simulation, WSC '99. ACM, New York, NY, USA, pp. 647–654.
Knol, A.B., Slottje, P., van der Sluijs, J.P., Lebret, E., 2010. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environmental Health 9, 19.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143.
Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5, 89–97.
Krueger, T., Page, T., Hubacek, K., Smith, L., Hiscock, K., 2012. The role of expert opinion in environmental modelling. Environmental Modelling and Software 36, 4–18.
Kuhnert, M., Voinov, A., Seppelt, R., 2006. Comparing raster map comparison algorithms for spatial modeling and analysis. Photogrammetric Engineering and Remote Sensing 71 (8), 975–984.
Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrology and Earth System Sciences 11 (4), 1267–1277.
Lane, S.N., 2007. Assessment of rainfall–runoff models based upon wavelet analysis. Hydrological Processes 21 (5), 586–607.
Littlewood, I.G., Croke, B.F.W., 2008. Data time-step dependency of conceptual rainfall streamflow model parameters: an empirical study with implications for regionalisation. Hydrological Sciences Journal 53 (4), 685–695.
Liu, F., Ma, P., Yang, M., 2005. A validation methodology for AI simulation models. In: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, vol. 7, pp. 4083–4088.
Ljung, L., 1999. System Identification – Theory For the User, second ed. Prentice Hall, Upper Saddle River, NJ, USA.
Makler-Pick, V., Gal, G., Gorfine, M., Hipsey, M.R., Carmel, Y., 2011. Sensitivity analysis for complex ecological models – a new approach. Environmental Modelling and Software 26, 124–134.
Marsili-Libelli, S., Arrigucci, S., 2004. Circadian patterns recognition in ecosystems by wavelet filtering and fuzzy clustering. In: Pahl, C., Schmidt, S., Jakeman, A. (Eds.), Proceedings of 2004 International Congress on Complexity and Integrated Resources Management. iEMSs, Osnabrück, Germany.
Marsili-Libelli, S., Checchi, N., 2005. Identification of dynamic models for horizontal subsurface constructed wetlands. Ecological Modelling 187, 201–218.
Marsili-Libelli, S., Giusti, E., 2008. Water quality modelling for small river basins. Environmental Modelling and Software 23, 451–463.
Marsili-Libelli, S., Guerrizio, S., Checchi, N., 2003. Confidence regions of estimated parameters for ecological systems. Ecological Modelling 165, 127–146.
Marsili-Libelli, S., 1992. Parameter estimation of ecological models. Ecological Modelling 62, 233–258.
Marsili-Libelli, S., 2006. Control of SBR switching by fuzzy pattern recognition. Water Research 40, 1095–1107.
Matott, L.S., Babendreier, J.E., Purucker, S.T., 2009. Evaluating uncertainty in integrated environmental models: a review of concepts and tools. Water Resources Research 45. http://dx.doi.org/10.1029/2008wr007301.
Matott, L.S., Tolson, B.A., Asadzadeh, M., 2012. A benchmarking framework for simulation-based optimization of environmental models. Environmental Modelling and Software 35, 19–30.
Matthews, K.B., Rivington, M., Blackstock, K., McCrum, G., Buchan, K., Miller, D.G., 2011. Raising the bar? – the challenges of evaluating the outcomes of environmental modelling and software. Environmental Modelling and Software 26 (3), 247–257.
McIntosh, B., Alexandrov, S.G., Matthews, K., Mysiak, J., van Ittersum, M. (Eds.), 2011. Thematic issue on the assessment and evaluation of environmental models and software. Environmental Modelling and Software, vol. 26 (3), pp. 245–336.
Miles, J.C., Moore, C.J., Kotb, A.S.M., Jaberian-Hamedani, A., 2000. End user evaluation of engineering knowledge based systems. Civil Engineering and Environmental Systems 17 (4), 293–317.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 50 (3), 885–900.

Moussa, R., 2010. When monstrosity can be beautiful while normality can be ugly: assessing the performance of event-based flood models. Hydrological Sciences Journal 55 (6), 1074–1084.
Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Monthly Weather Review 115 (7), 1330–1338.
Murray–Darling Basin Commission, 2000. Groundwater Flow Modelling Guideline. Murray–Darling Basin Commission, Canberra. Project No. 125.
Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models part i – a discussion of principles. Journal of Hydrology 10 (3), 282–290.
Nguyen, T.G., de Kok, J.L., Titus, M.J., 2007. A new approach to testing an integrated water systems model using qualitative scenarios. Environmental Modelling and Software 22 (11), 1557–1571.
Nocerino, J.M., Schumacher, B.A., Dary, C.C., 2005. Role of sampling devices and laboratory subsampling methods in representative sampling strategies. Environmental Forensics 6 (14), 35–44.
Norton, J.P., 2008. Algebraic sensitivity analysis of environmental models. Environmental Modelling and Software 23 (8), 963–972.
Norton, J.P., 2009. An Introduction to Identification. Dover Publications, Mineola, NY (Reprinted from Academic Press, London and New York, 1986).
Nossent, J., Elsen, P., Bauwens, W., 2011. Sobol' sensitivity analysis of a complex environmental model. Environmental Modelling and Software 26 (12), 1515–1525.
Olesen, H.R., 2005. User's Guide to the Model Validation Kit, Research Notes. NERI No. 226. Ministry of the Environment, Denmark.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263 (5147), 641–646.
Page, T., Heathwaite, A.L., Thompson, L.J., Pope, L., Willows, R., 2012. Eliciting fuzzy distributions from experts for ranking conceptual risk model components. Environmental Modelling and Software 36, 19–34.
Parker, P., Letcher, R., Jakeman, A., Beck, M., Harris, G., Argent, R., Hare, M., Pahl-Wostl, C., Voinov, A., Janssen, M., 2002. Progress in integrated assessment and modelling. Environmental Modelling and Software 17 (3), 209–217.
Perrin, C., Andréassian, V., Michel, C., 2006a. Simple benchmark models as a basis for criteria of model efficiency. Archiv für Hydrobiologie Supplement 161/1–2, Large Rivers 17 (1–2), 221–244.
Perrin, C., Dilks, C., Bärlund, I., Payan, J.L., Andréassian, V., 2006b. Use of simple rainfall–runoff models as a baseline for the benchmarking of the hydrological component of complex catchment models. Archiv für Hydrobiologie Supplement 161/1–2, Large Rivers 17 (1–2), 75–96.
Petersen, B., Gernaey, K., Devisscher, M., Dochain, D., Vanrolleghem, P.A., 2003. A simplified method to assess structurally identifiable parameters in Monod-based activated sludge models. Water Research 3, 2893–2904.
Petersen, B., 2000. Calibration, Identifiability and Optimal Experimental Design of Activated Sludge Models. PhD thesis in Applied Biological Sciences, BIOMATH, University of Gent. URL: http://biomath.ugent.be/publications/download/.
Piuleac, C.G., Rodrigo, M.A., Cañizares, P., Curteanu, S., Sáez, C., 2010. Ten steps of electrolysis processes by using neural networks. Environmental Modelling and Software 25, 74–81.
Pohjanpalo, H., 1978. System identifiability based on the power series expansion of the solution. Mathematical Biosciences 41, 21–33.
Pontius, R., Millones, M., 2010. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32 (15), 4407–4429.
Pontius, R., 2004. Useful techniques of validation for spatially explicit land-change models. Ecological Modelling 179 (4), 445–461.
Poole, D., Raftery, A.E., 2000. Inference for deterministic simulation models: the Bayesian melding approach. Journal of the American Statistical Association 95 (452), 1244–1255.
Pushpalatha, R., Perrin, C., Le Moine, N., Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. Journal of Hydrology 420–421, 171–182.
Raick, C., Soetaert, K., Gregoire, M., 2006. Model complexity and performance: how far can we simplify? Progress in Oceanography 70 (1), 27–57.
Ratto, M., Castelletti, A., Pagano, A., 2012. Emulation techniques for the reduction and sensitivity analysis of complex environmental models. Environmental Modelling and Software 34, 1–116.
Ravalico, J.K., Dandy, G.C., Maier, H.R., 2010. Management Option Rank Equivalence (MORE) – a new method of sensitivity analysis for decision-making. Environmental Modelling and Software 26, 171–181.
Refsgaard, J.C., Henriksen, H.J., 2004. Modelling guidelines – terminology and guiding principles. Advances in Water Resources 27 (1), 71–82. http://dx.doi.org/10.1016/j.advwatres.2003.08.006.
Refsgaard, J.C., Henriksen, H.J., Harrar, W.G., Scholten, H., Kassahun, A., 2005. Quality assurance in model based water management – review of existing practice and outline of new approaches. Environmental Modelling and Software 20 (10), 1201–1215.
Reusser, D.E., Blume, T., Schaefli, B., Zehe, E., 2009. Analysing the temporal dynamics of model performance for hydrological models. Hydrology and Earth System Sciences 13 (7), 999–1018.
Risbey, J., Kandlikar, M., Patwardhan, A., 1996. Assessing integrated assessments. Climatic Change 34 (3), 369–395.
Robson, B.J., Hamilton, D.P., 2004. Three-dimensional modelling of a Microcystis bloom event in the Swan River estuary, Western Australia. Ecological Modelling 174 (1–2), 203–222.
Robson, B.J., Hamilton, D.P., Webster, I.T., Chan, T., 2008. Ten steps applied to development and evaluation of process-based biogeochemical models of estuaries. Environmental Modelling and Software 23 (4), 367–384.
Robson, B., Cherukuru, N., Brando, V., 2010. Using satellite-derived optical data to improve simulation of the 3D light field in a biogeochemical model. In: Swayne, D.A., Yang, W., Voinov, A.A., Rizzoli, A., Filatova, T. (Eds.), Proceedings of 2010 International Congress on Environmental Modelling and Software. iEMSs, Ottawa, Canada.
Rowan, J.S., Greig, S.J., Armstrong, C.T., Smith, D.C., Tierney, D., 2012. Development of a classification and decision-support tool for assessing lake hydromorphology. Environmental Modelling and Software 36, 86–98.
Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90 (3), 229–244.
Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. Environmental Modelling and Software 25 (12), 1508–1517.
Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity Analysis. In: Wiley Series in Probability and Statistics. Wiley.
Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resources Research 46, W10531. http://dx.doi.org/10.1029/2009WR008933.
Schwarz, G.E., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2), 461–464.
Seber, G.A.F., Wild, C.J., 1989. Nonlinear Regression. John Wiley & Sons, New York.
Seibert, J., 2001. On the need for benchmarks in hydrological modelling. Hydrological Processes 15 (6), 1063–1064.
Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrology and Earth System Sciences 16 (4), 1171–1189.
Seppelt, R., Richter, O., 2005. "It was an artefact not the result": a note on systems dynamic model development tools. Environmental Modelling and Software 20 (12), 1543–1548.
Seppelt, R., Richter, O., 2006. Corrigendum to "It was an artefact not the result: a note on systems dynamic (Environ. Model. Softw. 20 (2005) 1543–1548). Environmental Modelling and Software 21, 756–758.
Seppelt, R., Voinov, A.A., 2003. Optimization methodology for land use patterns—evaluation based on multiscale habitat pattern comparison. Ecological Modelling 168 (3), 217–231.
Seppelt, R., Müller, F., Schröder, B., Volk, M., 2009. Challenges of simulating complex environmental systems at the landscape scale: a controversial dialogue between two cups of espresso. Ecological Modelling 220 (24), 3481–3489.
Shahsavani, D., Grimvall, A., 2011. Variance-based sensitivity analysis of model outputs using surrogate models. Environmental Modelling and Software 26, 723–730.
Smiatek, G., Kuntsmann, H., Werhahn, J., 2012. Implementation and performance analysis of a high resolution coupled numerical weather and river runoff prediction model system for an Alpine catchment. Environmental Modelling and Software 38, 231–243.
Sorooshian, S., Dracup, J.A., 1980. Stochastic parameter estimation procedures for hydrologic rainfall–runoff models: correlated and heteroscedastic error cases. Water Resources Research 16 (2), 430–442.
Stow, C.A., Joliff, J., McGillicuddy, D.J., Doney, S.C., Allen, J.I., Friedrichs, M.A.M., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems 76 (1–2), 4–15.
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.
Strang, G., Nguyen, T., 1996. Wavelets and Filter Banks. Wellesley-Cambridge Press.
Thunis, P., Georgieva, E., Pederzoli, A., 2012. A tool to evaluate air quality model performances in regulatory applications. Environmental Modelling and Software 38, 220–230.
Tompa, D., Morton, J., Jernigan, E., 2000. Perceptually based image comparison. In: Proceedings of the IEEE International Conference on Image Processing, Vancouver, BC, Canada, 2000, pp. 489–492.
Torrence, C., Compo, A., 1998. A practical guide to wavelet analysis. Bulletin of the American Meteorological Society 79, 61–78.
USEPA, 2002. Guidance for Quality Assurance Project Plans for Modeling. EPA QA/G-5M. United States Environmental Protection Agency, Office of Environmental Information, Washington DC.
USEPA, 2009. Guidance on the Development, Evaluation, and Application of Environmental Models. EPA/100/K-09/003. Office of the Science Advisor, Council for Regulatory Environmental Modeling. United States Environmental Protection Agency.
van der Sluijs, J., Janssen, P., Petersen, A., Kloprogge, P., Risbey, J., Tuinstra, W., Ravetz, J., 2004. RIVM/MNP Guidance for Uncertainty Assessment and Communication: Tool Catalogue for Uncertainty Assessment. Utrecht University. Retrieved 27th April 2011, from: http://www.nusap.net/sections.php.
van der Sluijs, J.P., Cray, M., Funtowicz, S., Kloprogge, P., Risbey, J., 2005. Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. Risk Analysis 25 (2), 481–492.
Vanrolleghem, P.A., Keesman, K.J., 1996. Identification of biodegradation models under model and data uncertainty. Water Science and Technology 33, 91–105.
Voinov, A., Bousquet, F., 2010. Modelling with stakeholders. Environmental Modelling and Software 25, 1268–1281.

Vrana, I., Vaníček, J., Kovář, P., Brožek, J., Alya, S., 2012. A group agreement-based approach for decision making in environmental issues. Environmental Modelling and Software 36, 99–110.
Vrugt, J., ter Braak, C., Gupta, H., Robinson, B., 2009. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stochastic Environmental Research and Risk Assessment 23 (7), 1011–1026. http://dx.doi.org/10.1007/s00477-008-0274-y.
Wagener, T., Kollat, J., 2007. Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo analysis toolbox. Environmental Modelling and Software 22 (7), 1021–1033.
Wagener, T., Boyle, D., Lees, M., Wheater, H., Gupta, H., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrology and Earth System Sciences 5 (1), 13–26.
Wagener, T., Mcintyre, N., Lees, M.J., Wheater, H.S., Gupta, H.V., 2003. Towards reduced uncertainty in conceptual rainfall–runoff modelling: dynamic identifiability analysis. Hydrological Processes 17 (2), 455–476.
Walter, E., Pronzato, L., 1997. Identification of Parametric Models from Experimental Data. In: Communications and Control Engineering Series. Springer, London.
Walter, E., 1982. Identifiability of State Space Models, with Applications to Transformation Systems. Springer, Berlin and New York, 202 pp.
Wealands, S., Grayson, R., Walker, J., 2005. Quantitative comparison of spatial fields for hydrological model assessment – some promising approaches. Advances in Water Resources 28 (1), 15–32.
Weijs, S.V., Schoups, G., van de Giesen, N., 2010. Why hydrological predictions should be evaluated using information theory. Hydrology and Earth System Sciences 14 (12), 2545–2558.
Willmott, C.J., 1981. On the validation of models. Physical Geography 2 (2), 184–194.
Yang, J., 2011. Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. Environmental Modelling and Software 26 (4), 444–457.
Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resources Research 44 (9), W09417. http://dx.doi.org/10.1029/2007wr006716.
Young, P.C., Jakeman, A.J., McMurtrie, R.E., 1980. An instrumental variable method for model order identification. Automatica 16, 281–294.
Young, P.C., 2011. Recursive Estimation and Time-series Analysis: an Introduction for the Student and Practitioner, second ed. Springer, New York.