Journal of Hydrology 480 (2013) 33–45

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

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

Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments

Axel Ritter [a], Rafael Muñoz-Carpena [b,*]

[a] Departamento de Ingeniería, Producción y Economía Agraria, Universidad de La Laguna, Ctra. Geneto, 2, 38200 La Laguna, Spain
[b] Agricultural and Biological Engineering, University of Florida, 287 Frazier Rogers Hall, PO Box 110570 Gainesville, FL 32611-0570, USA

ARTICLE INFO

Article history:
Received 26 September 2012
Received in revised form 28 November 2012
Accepted 3 December 2012
Available online 12 December 2012
This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Rao S. Govindaraju, Associate Editor

Keywords:
Block bootstrapping
Coefficient of efficiency
Hypothesis testing
Model prediction error

SUMMARY

Success in the use of computer models for simulating environmental variables and processes requires objective model calibration and verification procedures. Several methods for quantifying the goodness-of-fit of observations against model-calculated values have been proposed but none of them is free of limitations and are often ambiguous. When a single indicator is used it may lead to incorrect verification of the model. Instead, a combination of graphical results, absolute value error statistics (i.e. root mean square error), and normalized goodness-of-fit statistics (i.e. Nash–Sutcliffe Efficiency coefficient, *NSE*) is currently recommended. Interpretation of *NSE* values is often subjective, and may be biased by the magnitude and number of data points, data outliers and repeated data. The statistical significance of the performance statistics is an aspect generally ignored that helps in reducing subjectivity in the proper interpretation of the model performance. In this work, approximated probability distributions for two common indicators (*NSE* and root mean square error) are derived with bootstrapping (block bootstrapping when dealing with time series), followed by bias corrected and accelerated calculation of confidence intervals. Hypothesis testing of the indicators exceeding threshold values is proposed in a unified framework for statistically accepting or rejecting the model performance. It is illustrated how model performance is not linearly related with *NSE*, which is critical for its proper interpretation. Additionally, the sensitivity of the indicators to model bias, outliers and repeated data is evaluated. The potential of the difference between root mean square error and mean absolute error for detecting outliers is explored, showing that this may be considered a necessary but not a sufficient condition of outlier presence. The usefulness of the approach for the evaluation of model performance is illustrated with case studies including those with similar goodness-of-fit indicators but distinct statistical interpretation, and others to analyze the effects of outliers, model bias and repeated data. This work does not intend to dictate rules on model goodness-of-fit assessment. It aims to provide modelers with improved, less subjective and practical model evaluation guidance and tools.

1. Introduction

Most mathematical models used for calculating variables or simulating processes in hydrological and other environmental sciences must be previously evaluated with techniques that allow for their performance assessment. Generally, model performance is judged by comparing the calculated values and the corresponding measured or numerical benchmark data. Descriptions of various goodness-of-fit indices or indicators including their advantages and shortcomings, and rigorous discussions on the suitability of each index can be found elsewhere (Dawson et al., 2007; Harmel et al., 2010; Jain and Sudheer, 2008; Krause et al., 2005; Legates and McCabe, 1999; Loague and Green, 1991; Moriasi et al., 2007; Reusser et al., 2009; Santhi et al., 2001; Van Liew et al., 2003; Willmott, 1981). Among these indicators, the Nash and Sutcliffe (1970) coefficient of efficiency has received considerable attention in hydrological modeling (Clarke, 2008; Garrick et al., 1978; Gupta et al., 2009; Gupta and Kling, 2011; Houghton-Carr, 1999; McCuen et al., 2006; Moussa, 2010), although it is not commonly used in other fields of environmental science (Schaefli and Gupta, 2007). Yet, there is no general agreement on a standard procedure for evaluating model performance, while it is widely accepted that it should be approached in a multi-objective sense (e.g. Boyle et al., 2000; Gupta et al., 1998; Wagener, 2003; Willems, 2009). Among the various recommendations that can be found in the literature, there are three main points of agreement: (i) the need for a standard procedure for the evaluation of mathematical models (Moriasi et al., 2007), (ii) the inadequate use of correlation-based indices (such as the coefficient of determination, R^2) for quantifying model performance (Legates and Davis, 1997) and (iii) that

* Corresponding author. Tel.: +1 352 392 1864 x287; fax: +1 352 392 4092.
E-mail addresses: aritter@ull.es (A. Ritter), carpena@ufl.edu (R. Muñoz-Carpena).

0022-1694/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2012.12.004

model performance assessment should include at least one absolute value error indicator (in the variable units), one dimensionless index (or indicator of the relative error) for quantifying the goodness-of-fit, and a graphical representation of the relationship between model estimates and observations (Biondi et al., 2012; James and Burgess, 1982; Legates and McCabe, 1999; Loague and Green, 1991; Pushpalatha et al., 2012). However, interpretation of the values computed for the various goodness-of-fit indicators is sometimes controversial. This is because, first, the information representing each index has different meanings; and second, there are no standard criteria for the range of values that indicate when model performance is acceptable, good or very good. As a result, despite the need for agreed and standardized validation protocols in hydrological modeling (Biondi et al., 2012), interpretation of goodness-of-fit indicators may be always considered subjective. Therefore, there is a pressing need to provide modelers with improved and practical model evaluation guidance and tools for helping in the interpretation of those goodness-of-fit measures.

Another important and non-trivial issue is the statistical significance of the computed value of goodness-of-fit indicators, an aspect generally ignored. In a broader sense, Clark et al. (2011a) addressed hypothesis testing for hydrological modeling by promoting a multiple-hypothesis approach for improving model representations of hydrological processes, and for handling uncertainties arising from model structural ambiguities and data errors, but no specific consideration was given to the statistical testing of model performance indicators. These indicators are calculated normally as a unique value from the set (i.e., a sample) of deviations between the values calculated by the model and the corresponding observations. However, as with any random variable, the index value obtained from a sample may differ from the true value and an underlying probability distribution may exist (McCuen et al., 2006). For example, the coefficient of determination (R^2) quantifies the degree of collinearity between observations and model-calculated values. One advantage of R^2 is that its statistical distribution is well defined and therefore its statistical significance may be easily assessed (Legates and McCabe, 1999). This is not the case for the most recommended model goodness-of-fit indicators (Legates and McCabe, 1999). McCuen et al. (2006) proposed approximate equations for the probability distribution of the Nash and Sutcliffe (1970) coefficient of efficiency, but the basis for such equations and their wider applicability was not addressed. Since the probability distribution of goodness-of-fit indicators may be unknown or does not conform to a closed form, several authors (Legates and McCabe, 1999; Willmott et al., 1985) suggested the use of bootstrapping (Efron, 1981a,b; Efron and Tibshirani, 1993), a resampling method for statistical inference, to obtain a distribution of the performance index from which confidence intervals can be developed. Although the method can be computationally expensive for large datasets, the computational power of current computers makes it feasible to use bootstrapping for selected goodness-of-fit indicators commonly used in hydrological modeling.

The objective of this work is to advance in the performance evaluation of deterministic hydrological and other environmental models in a statistically rigorous way by developing a practical and unified framework based on the combination of a reduced number of evaluation techniques: (i) selected performance indicators; (ii) criteria for the goodness-of-fit rating; (iii) their statistical significance using a bootstrap method and hypothesis testing of the indicators exceeding threshold values; and (iv) evaluation of other data effects. New contributions of this work are the integration of goodness-of-fit criteria and their statistical significance into a unified framework that includes the implementation of bootstrapping for obtaining the probability distribution of the performance indicators, including the use of a specific bootstrap method when dealing with time series data; development of a straightforward method for statistically accepting or rejecting model performance; presenting the non-linear relationship between a performance index and model prediction error; and the analysis of the effects of repeated data values on the indicator value. To facilitate a wider application of the proposed evaluation framework, a public-domain computer application (presented in the Appendix) for the evaluation of generic observed vs. predicted datasets was developed and tested with a wide range of published experimental and modeling results.

2. Materials and methods

2.1. Selection of model evaluation indicators and acceptable model error

The selection of the length and information content of the experimental data series used in the model evaluation is a critical step since it has strong influence on the model performance. Thus, methods for identifying critical time periods in the observations, which contain most of the information for parameter identification, can provide useful guidance when selecting the evaluation data series (Bárdossy and Singh, 2008; Singh and Bárdossy, 2012). For addressing model performance evaluation the use of qualitative and quantitative criteria is recommended, because both approaches allow for capturing distinct aspects of model performance (e.g. Biondi et al., 2012; Clark et al., 2011b; Kavetski and Clark, 2011; McMillan et al., 2011; Pushpalatha et al., 2012; Seibert and McDonnell, 2002; Young and Ratto, 2009). Based on previous recommendations (James and Burgess, 1982; Legates and McCabe, 1999; Loague and Green, 1991), a combination of three evaluation tools was adopted as a starting point in this work. Firstly, in order to evaluate model performance, a plot of observations against the calculated values illustrates the degree to which the points match the identity line (denoted as the 1:1 line). This scatter plot allows for a visual inspection of model performance, such that the higher the agreement between calculated and observed values is, the more the scatters tend to concentrate close to the 1:1 line. Another interesting aspect of this graphical method is its ability to show (i) if model performance is homogeneous along the prediction range or depends on the magnitude of the calculated values; (ii) other unexpected relationships between the two data sets: non-linear or a linear correlation different to the identity line; (iii) the presence of potential outliers; and (iv) appreciable model bias. Secondly, for quantifying the prediction error in terms of the units of the variable calculated by the model, the root mean square error (*RMSE*) is selected. This indicator is frequently used and its definition is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}} \tag{1}$$

where O_i and P_i represent the sample (of size N) containing the observations and the model estimates, respectively. It ranges from 0 to ∞, where $RMSE = 0$ indicates a perfect fit. Thirdly, as a dimensionless goodness-of-fit indicator, the Nash and Sutcliffe (1970) coefficient of efficiency is selected. This is calculated as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} = 1 - \left(\frac{RMSE}{SD}\right)^2 \tag{2}$$

where $\overline{O}$ is the mean of the observed values and SD represents the standard deviation of the observations. Thus, *NSE* represents the complement to unity of the ratio between the mean square error of observed vs. predicted values and the variance of the observations. The coefficient of efficiency takes values $-\infty \leqslant NSE \leqslant 1$.

A *NSE* = 1 indicates a perfect fit, while a $NSE \leqslant 0$ suggests that the mean of the observed values is a better predictor than the evaluated model itself. Gupta and Kling (2011) showed that for model optimization *NSE* typically varies within a finite range. Moreover, assuming a reasonably conceptualized model structure, they stated that negative *NSE* values should not generally happen unless there are severe errors in the input or output data.

The *NSE* is a widely used indicator in hydrology due to its flexibility to be applied to various types of mathematical models (Gupta and Kling, 2011; McCuen et al., 2006). According to McCuen et al. (2006), the *NSE* may be a useful goodness-of-fit indicator, but its limitations should be taken into account, namely its sensitivity to bias in model predictions and the possible effect on *NSE* of outliers present in the series $\{O_i, P_i\}$. Because of this, its suitability has been subject to discussion for many years (Ehret and Zehe, 2011; Gupta et al., 2009; Jain and Sudheer, 2008; Krause et al., 2005; Legates and McCabe, 1999; Martinec and Rango, 1989; McCuen and Snyder, 1975; McCuen et al., 2006; Schaefli and Gupta, 2007). Several authors have proposed modifications to the *NSE* (e.g. Krause et al., 2005; Le Moine, 2008; Oudin et al., 2006) or alternative indicators (e.g. Criss and Winston, 2008; Krause et al., 2005). Some modified forms of the *NSE* are based on transforming observations and model estimates (using root squared, log or inverse transformed series (Le Moine, 2008; Oudin et al., 2006): $\{\sqrt{O_i}, \sqrt{P_i}\}$, $\{\ln(O_i+\varepsilon), \ln(P_i+\varepsilon)\}$ and $\left\{\frac{1}{O_i+\varepsilon}, \frac{1}{P_i+\varepsilon}\right\}$, respectively. Another main modification suggests replacing the mean of the observations as the baseline model by appropriate benchmark series such as the seasonal or climatological means (Garrick et al., 1978; Legates and McCabe, 1999; Martinec and Rango, 1989; Murphy, 1988; Seibert, 2001). In this respect, while Schaefli and Gupta (2007) emphasize the importance of selecting an appropriate benchmark for each particular case study, Legates and McCabe (2012) point out that choosing proper benchmarks is however not straightforward and is not likely to be globally applicable. Besides, since such benchmarks may depend on the type of hydrological regime or model application, the indicator interpretation could be difficult for inexperienced end-users, who may simply want to know whether model performance can be rated as 'good', 'acceptable' or 'bad' (Pushpalatha et al., 2012). Legates and McCabe (2012) maintain the recommendation of using *NSE* and its modified form (i.e. using absolute instead of squared deviations in Eq. (2)) proposed by Legates and McCabe (1999). They argue that these are superior and preferable to many other statistics, because of intuitive interpretability and because these indicators have a fundamental meaning at zero.

While the *NSE* is the most widely used performance measure in hydrological modeling (Ewen, 2011; Guinot et al., 2011; Gupta et al., 2009; Pushpalatha et al., 2012), there are no globally accepted standards on the intervals to be used for the qualitative interpretation of *NSE*. Some authors (as indicated by Legates and McCabe, 1999) warn that goodness-of-fit indices are frequently misinterpreted because of the perception that they provide the same information as the coefficient of determination. Thus, while R^2 = 0.6 indicates that the model explains 60% of the variance in the observed data, a *NSE* = 0.6 has a totally different meaning, i.e., that the model mean squared error represents 40% of the observed variance.

To interpret *NSE* values, we propose to relate this indicator to the model prediction error it contains. Basically, model efficiency is considered satisfactory when prediction error in the units of the variable is "small". Additionally, model performance should account also for the width of the range covered by computed values (i.e., a model predicting with small error within a small range should not be better than a model yielding a larger error but within a wider range of the observations). Thus, for determining when the mean model error is "small" it can be compared to the variability of the observations. Satisfactory model efficiency can then be established depending on the number of times (n_t) that the observations variability is greater than the mean error. If the mean model error is represented by the *RMSE* and the variability of the observations is given by their standard deviation (*SD*), n_t is expressed as:

$$n_t = \frac{SD}{RMSE} - 1 \tag{3}$$

Combining Eqs. (2) and (3) the following relationship can be derived, which relates the *NSE* with our goodness-of-fit target variable, namely n_t:

$$NSE = 1 - \left(\frac{1}{n_t + 1}\right)^2 \tag{4}$$

Note that Eq. (4) shows that model performance does not change linearly with *NSE* (Fig. 1). For example, a change in *NSE* from 0.80 to 0.90 implies a change in the mean error of $n_t \approx 1$ where a *NSE* variation from 0.70 to 0.80 is just $n_t \approx 0.4$.

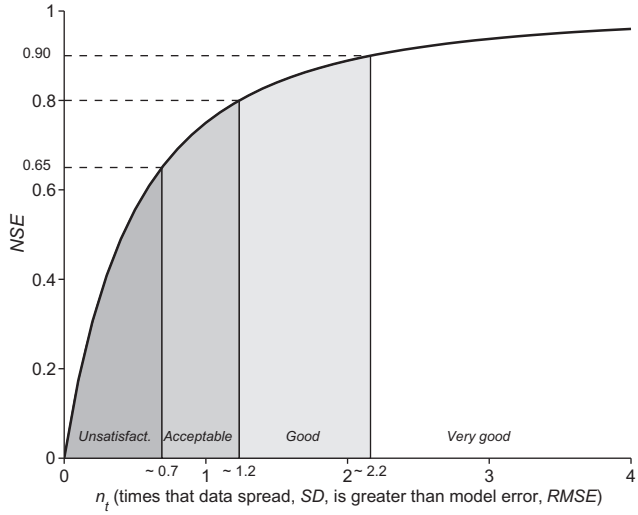

Fig. 1. Relationship between *NSE* and model mean error relative to the spread of the observations.

2.2. Model efficiency classes for hypothesis testing

Defining limits of acceptability for model predictions depends on model applications (Beven, 2006). Although subjective, in this work we propose four model performance classes based on n_t (Table 1) as a guidance on reference *NSE* ranges, which are denoted as *Unsatisfactory*, *Acceptable*, *Good* and *Very good*. The corresponding *NSE* limits were derived first based on a value of $NSE_{threshold}$ = 0.65, which has been reported in the literature as a lower limit of a valid goodness-of-fit (e.g. Moriasi et al., 2007). This is proposed here as the threshold value for an *Acceptable* model efficiency ($NSE \geqslant 0.65$), and according to Eq. (4) corresponds to $n_{t,threshold} = n_{t,0.65} = 0.69$. The next classification limits are established by considering mean error

Table 1
Criteria for the goodness-of-fit evaluation.

Performance rating	Model efficiency interpretation	n_t[a]	*NSE*
Very good	SD ⩾ 3.2 RMSE	⩾2.2	⩾0.90
Good	SD = 2.2 RMSE–3.2 RMSE	1.2–2.2	0.80–0.90
Acceptable	SD = 1.2 RMSE–2.2 RMSE	0.7–1.2	0.65–0.80
Unsatisfactory	SD < 1.7 RMSE	<0.7	<0.65

[a] n_t: Times that spread of observations (*SD*) is greater than mean model error (expressed as *RMSE*).

multiples from the lower limit (i.e. $2n_{t,0.65}$ and $3n_{t,0.65}$), which correspond approximately to *NSE* = 0.80 and *NSE* = 0.90, respectively (Eq. (4)). Fig. 1 and Table 1 summarize these four model efficiency classes and their relationship with *NSE*, *RMSE*, n_t and *SD* values. Other model efficiency thresholds for particular applications could be justified and used without affecting the value of the methods proposed herein. Those presented and evaluated here represent a first approximation to advance the adoption of the hypothesis significance testing in the evaluation of hydrological models.

2.3. Statistical significance of the selected indicators

The statistical significance of the selected performance indicators (*RMSE* and *NSE*) can be assigned if their probability distributions were available. This is not generally the case since such distributions are unknown or do not have a closed form -as is the case of the *RMSE*. Because the observed dataset is just one realization of the system behavior, obtaining random sub-samples of the observed data and then calculating the performance indicators for each of the pseudo-observed samples and the corresponding simulated values would allow for building such probability distributions. This, however, is not straightforward. The bootstrapping method (Efron, 1979), a Monte Carlo sampling technique, may be useful for approximating the indicators' probability distributions. Like other resampling methods, it consists basically in drawing repeated samples (i.e. resamples) from the available dataset and treating them as if they were the result of actual sampling from the unknown population. This allows for generating an approximate of the statistic's distribution from the available dataset by drawing a large number of resamples and calculating the value of the statistic for each of them. Each resample has the same number of elements as the available dataset, but it may include some of the original data points more than once, while some may not be included. Therefore, the resamples will randomly depart from the original available dataset, thus providing an approximation of the statistic's distribution.

Efron's original nonparametric bootstrap procedure assumes that data are independent and identically distributed. For weakly dependent stationary observations, i.e. many environmental time series, block bootstrapping may be applied, instead. Basically, this consists of dividing the data set into blocks that are approximately independent and then perform random sampling with replace-

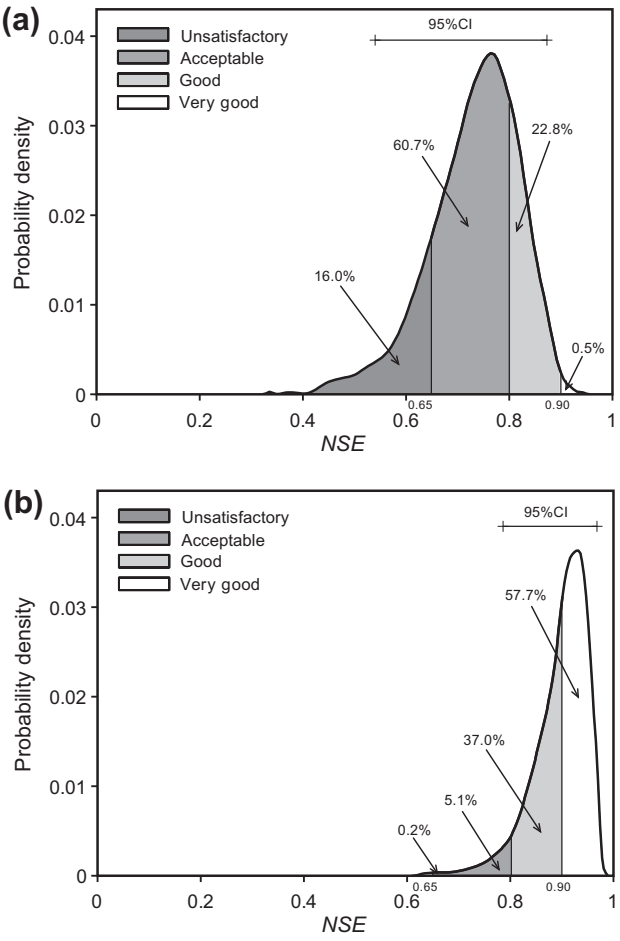

Fig. 2. *NSE* probability distribution obtained by bootstrapping and corresponding *NSE* statistical significance. Examples where the goodness-of-fit is considered not valid (a) and valid (b) at 0.10 significance level.

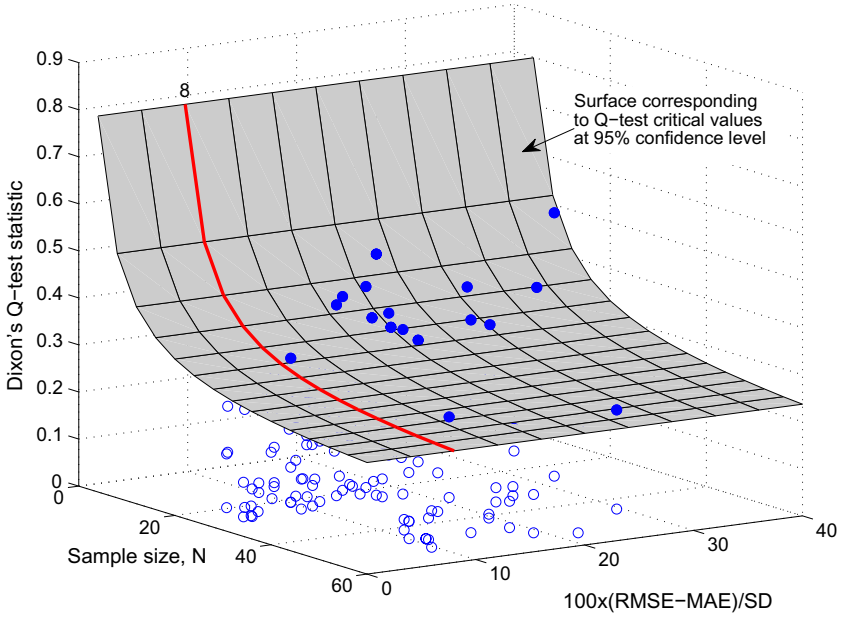

Fig. 3. Difference between *RMSE* and *MAE* relative to the observations standard deviation (*SD*) vs. Dixon's outlier Q-test (68 cases). Solid symbols indicate cases where the test detects an outlier candidate.

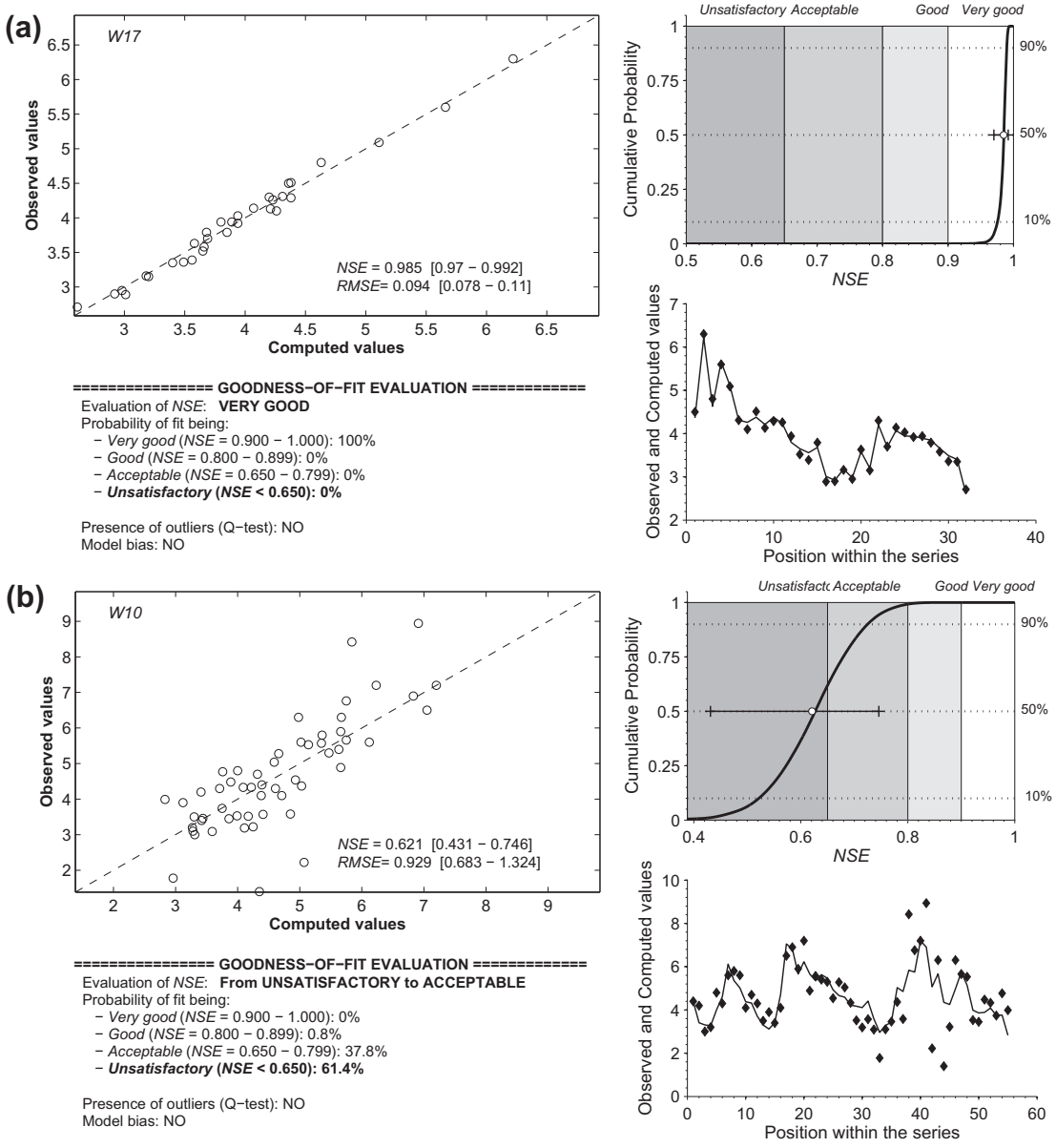

Fig. 4. Goodness-of-fit evaluation examples for two of the four performance classes proposed: (a) *Very good*, and (b) *Unsatisfactory*. The simulated variable is groundwater nitrogen-nitrate concentration (mg L^{-1}) as described in Ritter et al. (2007). Outputs were obtained using the computer tool presented in the Appendix.

ment on these blocks rather than on the individual observations. Thereby the original data structure is aimed to be preserved within each a block (Carlstein, 1986; Künsch, 1989). Several variants of block bootstrap have been proposed: the non-overlapping block bootstrap (Carlstein, 1986); the moving block bootstrap (Künsch, 1989; Liu and Singh, 1992); the circular block bootstrap (Politis and Romano, 1992), and the stationary bootstrap (Politis and Romano, 1994), among others. For strongly dependent time series, Kapetanios and Papailias (2011) introduced a procedure, which basically consists in transforming the original observations into a weak dependent data set, obtaining block bootstrap resamples from these data and finally proceeding with a back-transformation. The performance of block bootstrap depends however on the choice of the block size length used. Politis and White (2004) and Patton et al. (2009) derived a data-dependent method for the automatic selection of the optimal block length for both the circular and the stationary block bootstrap. According to Politis and White (2004), the stationary bootstrap is less sensitive to block size misspecification and estimating the optimal block size for this method seems to be an easier problem than in the circular or in the moving block bootstrap approaches.

In this work the vector of observed values (O_i) is a taken for generating M bootstrap resamples (of equal size N) using either Efron and Tibshirani (1993) or Politis and Romano (1994) methods. Then, taking the corresponding model-predicted values (P_i) the *RMSE* and *NSE* are calculated for each resample. This allows us to construct an empirical probability distribution of the *RMSE* and *NSE* statistics. Depending on the number of resamples M (here we used $M = 2000$), the procedure may be computationally expensive. Once the probability distribution functions of the *RMSE* and *NSE* indicators are approximated, their statistical significance may be assessed based on the 95% confidence interval. This is computed with the bias corrected and accelerated method, *Bca* (DiCiccio and Efron, 1996), which adjusts for both bias and skewness in the bootstrap distribution. Goodness-of-fit reliability can be also addressed in more detail by computing the probability of given

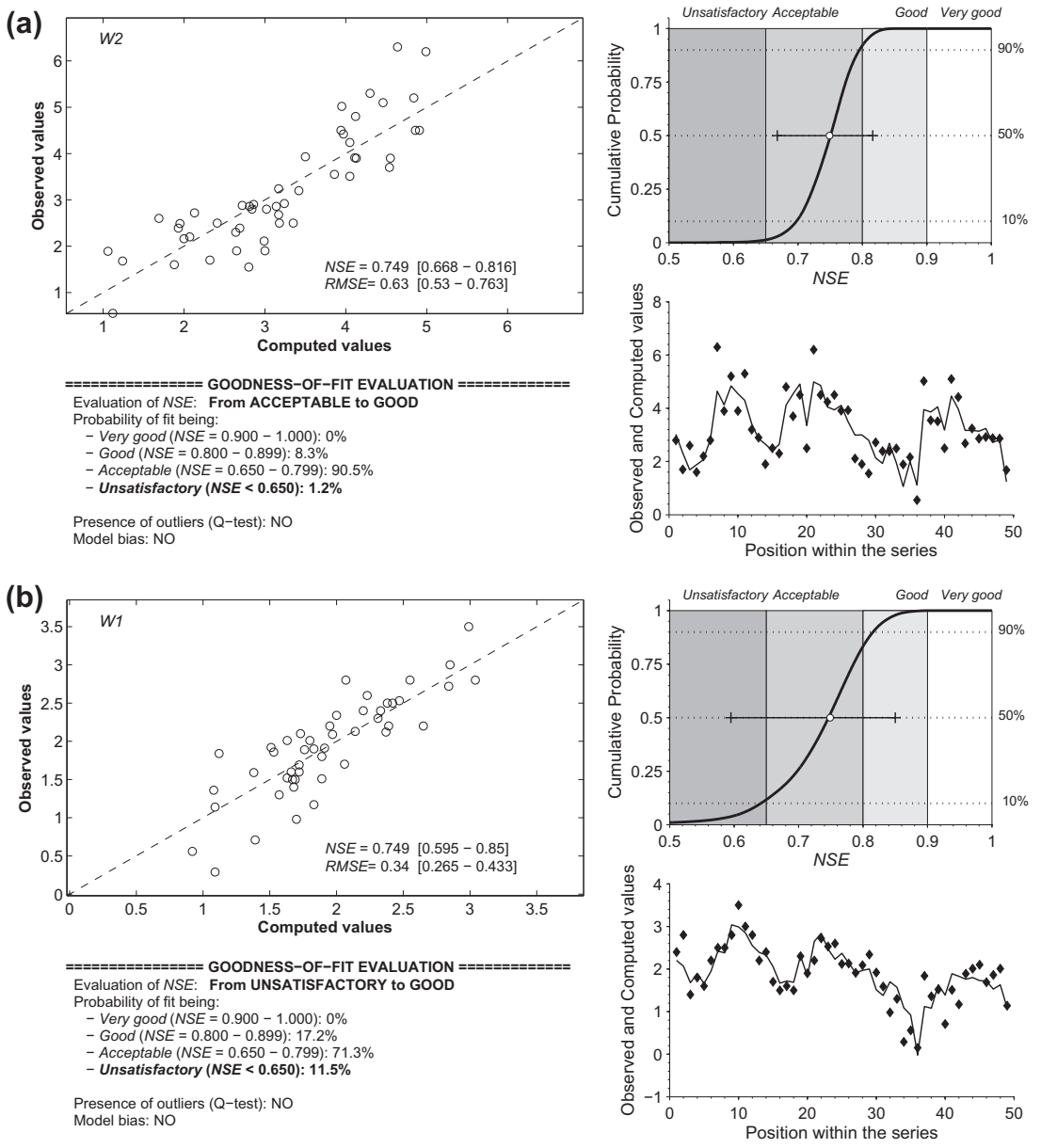

Fig. 5. Goodness-of-fit evaluation of two cases with equal *NSE* but different statistical significance. Valid (a) and not valid (b) model performance at 0.10 significance level. The simulated variable is groundwater nitrogen-nitrate concentration (mg L^{-1}) as described in Ritter et al. (2007). Outputs were obtained using the computer tool presented in the Appendix.

NSE values (e.g., the probability that the *NSE* is within certain acceptable limits, Table 1) from the estimated probability distribution function.

Additionally, once the approximate probability distribution is known, it is possible to test statistical hypotheses on the model goodness-of-fit. For example, here the null hypothesis (H0) represents that the median *NSE* is less than the threshold *NSE* value below which the goodness of fit is not acceptable ($NSE < NSE_{threshold}$) and the alternative hypothesis (H1) when it is acceptable ($NSE \geqslant NSE_{threshold}$). The null hypothesis is rejected and the result (i.e. goodness-of-fit is acceptable) is statistically significant when the *p*-value is less than a significance level α. The *p*-value represents here the probability of wrongly accepting the fit ($NSE \geqslant NSE_{threshold} = 0.65$), when it should be rejected (i.e., when H0 is true). Common significance level values are $\alpha = 0.1$, 0.05 or 0.01 but the choice of α should be based on the research context, i.e. how strong the evidence needs to be for accepting or rejecting H0. As a starting point, here we suggest adopting the least restrictive significance level of $\alpha = 0.10$.

Fig. 2 illustrates two examples of *NSE* probability distributions obtained by bootstrapping, where shaded areas correspond to the probability of the goodness-of-fit being within the proposed performance classes (*Unsatisfactory*, *Acceptable*, *Good or Very good*). In Fig. 2a the 95% confidence interval of *NSE* [0.540–0.873] indicates that for this application the goodness-of-fit rating ranges from *Unsatisfactory to Good.* Focusing on the model efficiency classes proposed in Table 1, it follows that the fit is considered valid with a probability of only 84%, where the probability of *Very good* ($NSE \geqslant 0.9$), *Good* (0.899–0.75) and *Acceptable* (0.65–0.749) is 0.5%, 22.8% and 60.7%, respectively. Similarly, Fig. 2b illustrates an example of an *Acceptable* to *Very good* fit, (95% confidence interval [0.786–0.968]). Here, the probability of the model fit being considered *Very good*, *Good* and *Acceptable* are 57.7%, 37.0% and 5.1%, respectively. Fig. 2 also indicates that, when adopting a significance

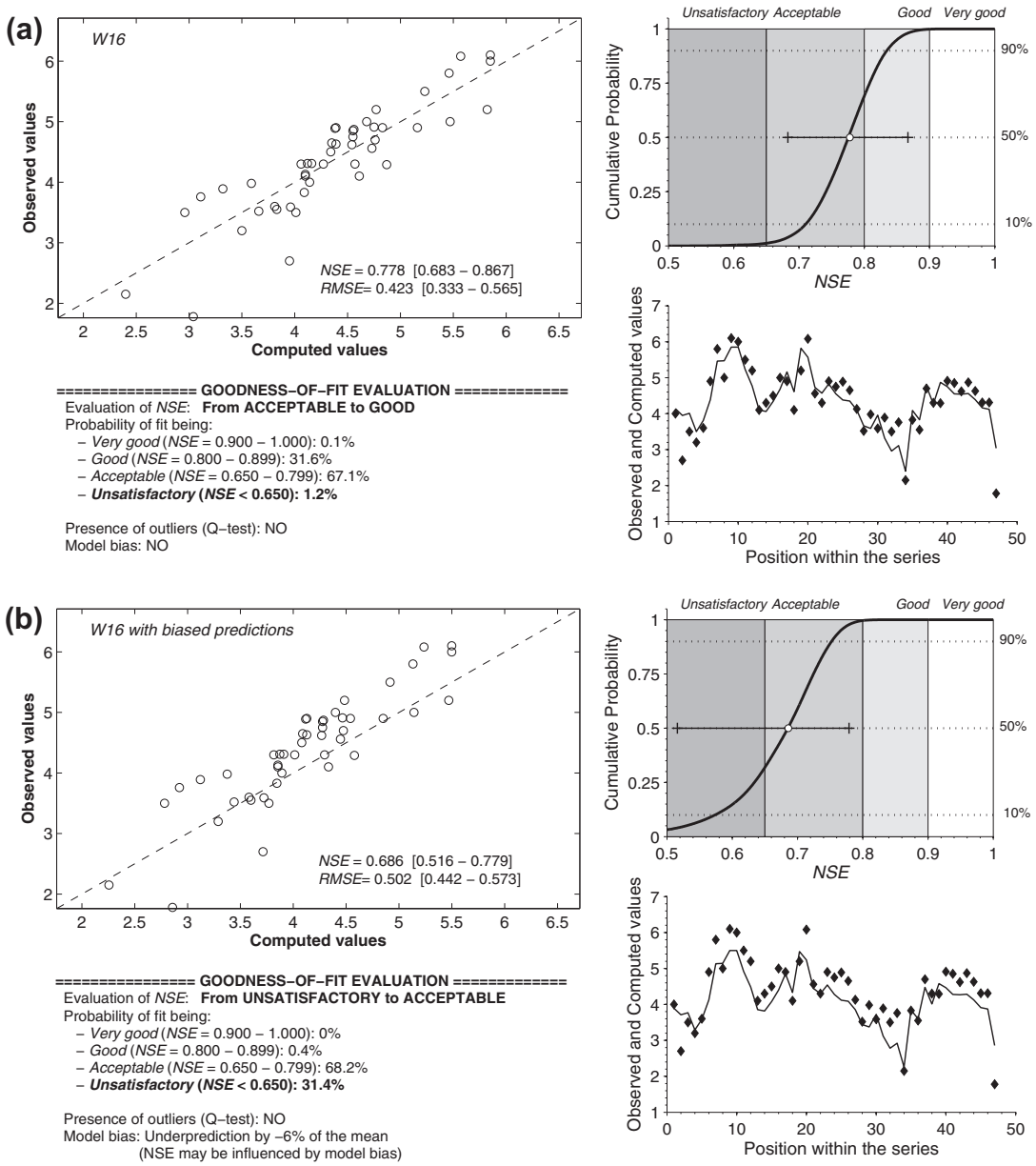

Fig. 6. Effect of bias in the calculated values on the goodness-of-fit assessment by reducing 6% the estimates of groundwater nitrogen-nitrate concentrations (mg L^{-1}) of example taken from Ritter et al. (2007). Outputs were obtained using the computer tool presented in the Appendix.

level of $\alpha = 0.10$, only the second of the model evaluations (Fig. 2b) would be considered valid (p-value = 0.002 < 0.10; reject H0). These examples illustrate the useful information provided by the proposed significance testing method.

2.4. Addressing the limitations of the coefficient of efficiency

2.4.1. Effect of biased predictions on NSE

In addition to obtaining a model with good predictions (i.e., a small error), it is also desirable to have an unbiased model (i.e., the sum of the errors is equal to zero). Due to systematic errors, the calculations of non-regression models can be biased and as reported by McCuen et al. (2006), the *NSE* value can be affected by biased model predictions. These authors proposed that the *NSE* must be accompanied by a quantification of the relative bias, defined as the ratio between the average deviations between calculated and observed values to the mean of observed values $\bar{O}$ (McCuen et al., 2006),

$$Relative\ bias = \frac{\frac{1}{N}\sum_{i=1}^{N}(P_i^* - O_i^*)}{\bar{O}^*}100; \quad \bar{O}^* > 0 \tag{5}$$

where O_i^* and P_i^* are the series of observations and estimates shifted into the positive range as needed. A relative bias (in absolute value) exceeding a given percentage value indicates that the model can be over- or underpredicting the observed data (McCuen et al., 2006). Such relative bias threshold value should be established by the modeler based on the quality of the data used in the model. In this work we follow McCuen et al. (2006) such that a relative bias greater than 5% in absolute value is considered significant. The presence of negative values in the series of observed and model calculated values may also affect the correct estimation of bias. Therefore, when negative values are present, the $\{O_i, P_i\}$ series are shifted into the positive range by an amount equal to the most negative value. Thus, a model that overestimates consistently the observations by >5% has a positive bias, while a consistent underestimation by >5% implies a negative bias. Note that seeking for an unbiased

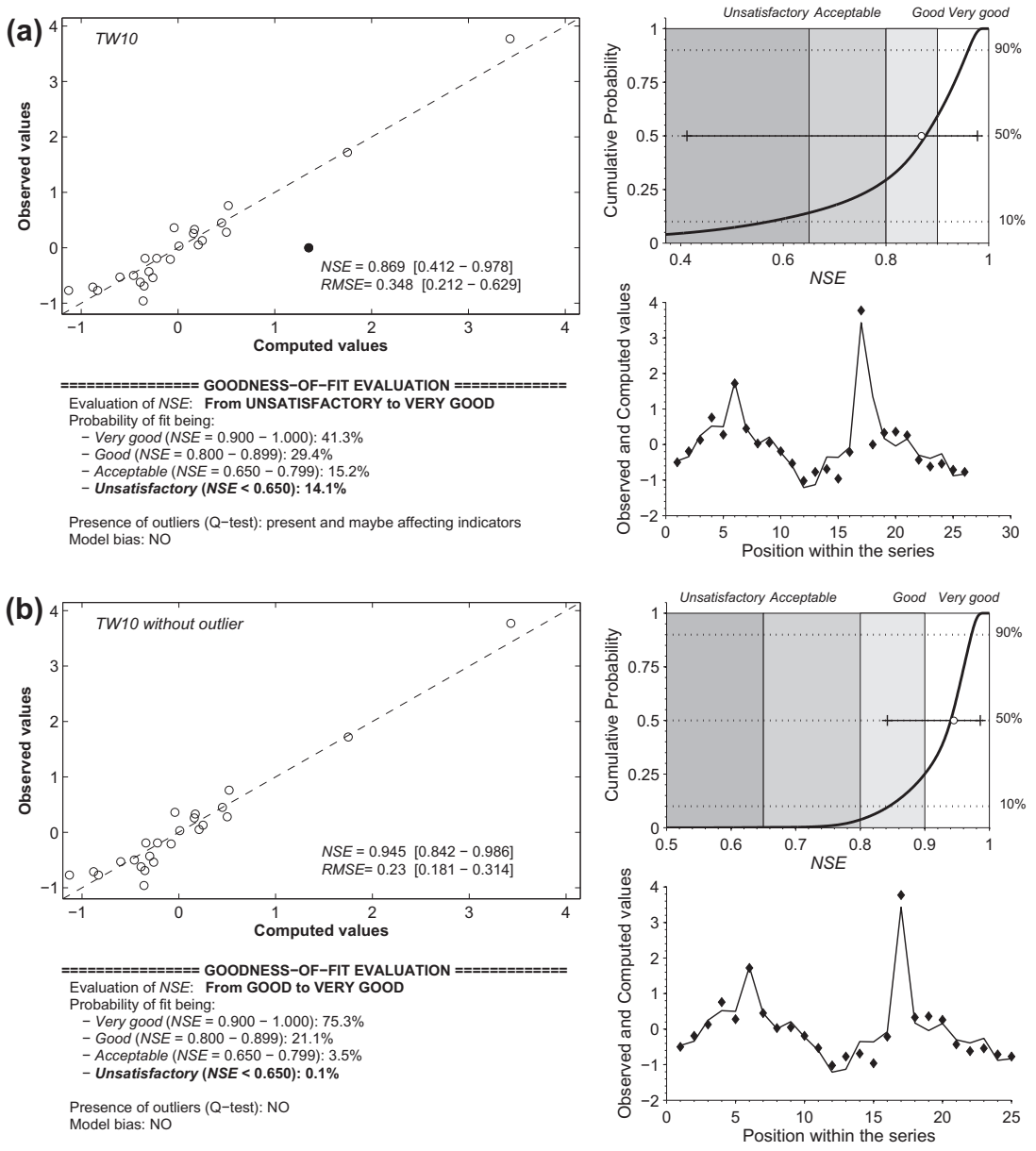

Fig. 7. Effect of the presence of outliers (solid symbol) on the goodness-of-fit evaluation. Evaluation for cases with (a) and without (b) the outlier candidate. The simulated variable is groundwater total phosphorus concentration (mg L^{-1}) as described in Muñoz-Carpena et al. (2005). Outputs were obtained using the computer tool presented in the Appendix.

model may increase prediction error and so the need to compromise between these two criteria (zero-bias model and minimum prediction error) adds an element of subjectivity in the model evaluation process that is not desirable but probably necessary (McCuen, 2003).

2.4.2. Effect of outliers on goodness-of-fit indices

Since both the *NSE* and the *RMSE* are calculated by squaring the deviations between observations and model-calculated values, the presence of an outlier can introduce a disproportionate effect on the value of goodness-of-fit statistics. The degree to which the *RMSE* is higher than the mean absolute error (*MAE*) has been proposed as an indicator of the presence of outliers (Legates and McCabe, 1999). However, there is no quantitative criterion available for evaluating when the difference between *RMSE* and *MAE* indicates outlier presence. To explore this, the value of the error difference (*RMSE–MAE*) expressed as a percentage of the variability of the observations (*SD*) was compared with an outlier detection method (Dixon's Q-test, Rorabacher (1991)) applied to the error vector (O_i–P_i). The comparison was performed on a set of 68 model evaluations obtained from Muñoz-Carpena et al. (2005) and Ritter et al. (2007) to warn when *RMSE* and *NSE* may be affected by the presence of outliers. Fig. 3 shows the degree to which the criterion (i.e. *RMSE* is higher than the *MAE*) can be taken as an indicator of the presence of outliers for this preliminary dataset. The solid symbols represent 17 out of the 68 cases containing an outlier, i.e. points above the surface in Fig. 3 with $(RMSE–MAE)/SD \geqslant 8$. However, this seems to be a necessary but not a sufficient condition, because 38 out of the remaining cases also exhibit a value $\geqslant 8$, but do not contain outliers (hollow symbols in Fig. 3). Based on this limited testing it was not possible to obtain a simple, yet robust criterion comparing squared and absolute errors as an indicator of outlier presence. Instead, the use of an

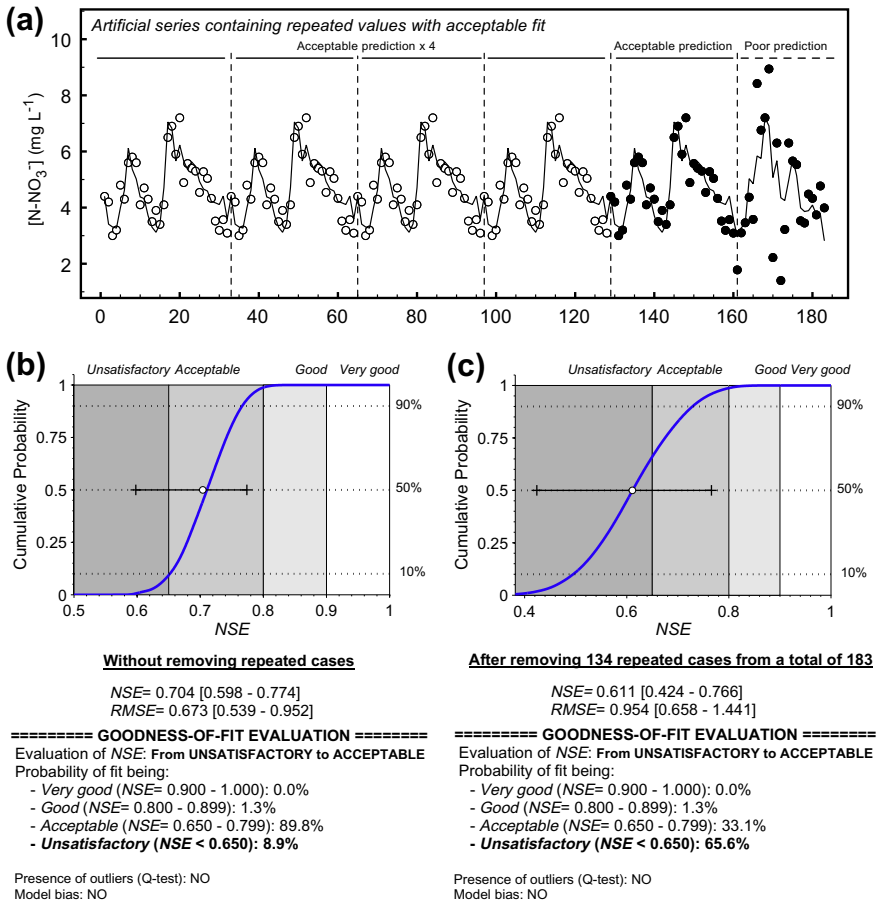

Fig. 8. Effect of repeated cases in the observed and predicted sets on the goodness-of-fit assessment. When repeated values improve the goodness-of-fit. Outputs were obtained using the computer tool presented in the Appendix.

established outlier detection test is recommended as a component of the model evaluation framework as illustrated in the examples presented below.

2.4.3. Effect of repeated cases in the observed and predicted sets on goodness-of-fit indices

Assessment of model performance takes into account the deviations between observations and calculated values sampled within a range. Given the same model input factors, it is generally expected that repeated measurements in the data series are associated with repeated model predictions. The number of repeated cases in the series may have an influence on the calculation of the goodness-of-fit indices. For a data set containing many times the same measurement and its corresponding prediction, if they yield a small deviation (O_i–P_i), then the model performance will be overestimated. By contrast, if they produce a large deviation, then the model performance will be underestimated. Therefore, a useful test (albeit cumbersome) to evaluate repeated values effects is to recalculate the goodness of fit after removal of repeated values and compare the results against those obtained with the complete data set. Next section includes also an example that shows the effect of repeated cases in the data set.

3. Illustrative case studies

In order to illustrate the application of the proposed model evaluation framework, example cases are provided below using a generic computer tool to facilitate the analyses. This incorporates the graphical and statistical evaluation techniques proposed herein, as is described in the Appendix. The procedure was first applied to results presented by Ritter et al. (2007). In this paper a dynamic factor model (see Zuur et al., 2003) was used for predicting time series of shallow groundwater nitrate–nitrogen concentration, [N—NO_3^-], in 18 monitoring wells distributed across a 4-ha agricultural field.

Fig. 4 illustrates the results obtained for two well nitrate concentrations, which yielded different model performance ratings. Fig. 4a shows a *Very good* (*NSE* = 0.985 [0.970–0.992]) prediction of groundwater [N—NO_3^-], and Fig. 4b corresponds to a case where the fit results span from *Unsatisfactory* to *Acceptable* (*NSE* = 0.621 [0.431–0.746]). The excellent model performance in the first case study may be also inferred from the plots in Fig. 4a, since they clearly show that computed values are very similar to the observations. In Fig. 4b, scatters follow the 1:1 line, but calculated values deviate from the observations positively and negatively along the prediction range. The evolution of observed and computed values in Fig. 4b indicates that model prediction deteriorates in the last period of the time series. The probability that the fit could be considered within a particular performance-rating category provides interesting information that enriches the evaluation of the performance of the dynamic factor model. The statistical inference for evaluating model performance is straightforward when focusing on the probability of the fit being *Unsatisfactory*. Model validity in Fig. 4a is significant since there is no probability of obtaining a *NSE* < 0.65. By contrast, Fig. 4b illustrates how in this case the probability that *NSE* < 0.65 is as high as 61.4%. For delimiting the rejection region we have proposed in Section 2.3 a cut-off value of 10% ($\alpha = 0.10$), which corresponds here to the lower dashed line in the upper-right panel of Fig. 4a and b. This means that for Fig. 4a, H0

can be rejected and the model goodness-of-fit is deemed statistically valid, but not for the case in Fig. 4b where H0 cannot be rejected (p-value = 0.614 > α = 0.10) and the model fitting is not considered acceptable.

Fig. 5 presents two additional and particular cases that illustrate how the goodness-of-fit evaluation based only on a single indicator value (*NSE*) lead to different conclusions if statistical significance is considered. In these examples both model fits (Fig. 5a and b) yield *NSE* = 0.749 and suggest equal model efficiency. However, based on the evaluation procedure proposed here, it follows that Fig. 5a and b represent two distinct cases where prediction of groundwater [$N{-}NO_3^-$] in well *W2* (Fig. 5a) varies from *Acceptable* to *Good* while for well *W1* (Fig. 5b) the model is rated as from *Unsatisfactory* to *Good*. The statistical significance test indicates that although both cases show a *NSE* = 0.749, the fit in well *W1* should be rejected (p-value = 0.115), while that in well *W2* can be accepted (p-value = 0.012). These two examples represent the effect of the model performance within a wider observed range. While the example in Fig. 5a yields a *RMSE* almost twice as that in Fig. 5b (0.63 vs. 0.34 mg/L), the observation range (i.e. *SD*) in Fig. 5a (lower right graph) is larger than that of Fig. 5b ([0 to ~6] vs. [0 to ~3.5]). The larger *SD* in Fig. 5a compensates the differences in *RMSE* (Eq. (2)), yielding the same *NSE* = 0.749.

In order to illustrate the effect of bias on the goodness-of-fit assessment, an additional example is included (well *W16*, Fig. 6a) where the goodness-of-fit is likely to vary from *Acceptable to Good* (*NSE* = 0.778 [0.683–0.867]). The computed values were then systematically decreased by 6% and the evaluation results compared (Fig. 6b). The bias in model predictions results in deterioration of goodness-of-fit where the model is now rated as *Unsatisfactory* to *Good*, the *NSE* reduces to 0.686 [0.516–0.779] and the fit deemed not valid statistically with a p-value = 0.314 (compared with the p-value = 0.012 in Fig. 6a).

The effect of the presence of outliers is illustrated with an example taken from Muñoz-Carpena et al. (2005), where a dynamic factor model was used for predicting time series of shallow groundwater concentration of total phosphorus, [TP] (Fig. 7). In this case, the set of observed and computed groundwater [TP] contains an outlier represented by point {1.35, 0.00} (Fig. 7a). The graphical comparison of observed and calculated values shows here the location of the outlier in the data series and illustrates the magnitude of the deviation of this particular point with respect to the other values in the data sets. Elimination of this point from the series (Fig. 7b), yields an improvement in goodness-of fit indicators (*NSE* = 0.945, *RMSE* = 0.230 mg L^{-1} vs. *NSE* = 0.869, *RMSE* = 0.348 mg L^{-1}), an increase in the probability of the fit being *Very good* (from 41.3% to 75.3%), and a considerable decrease in the likelihood of the fit being *Unsatisfactory* (from 14.1% to 0.1%), such that the fit may be now considered valid at the 0.10 significance level. Notice that the information gained here should not lead to the removal of outlier candidates from the dataset (unless there is a strong justification based on knowledge of the experimental data set), rather the analysis offers insights on potential causes of the model deviation from the measured results.

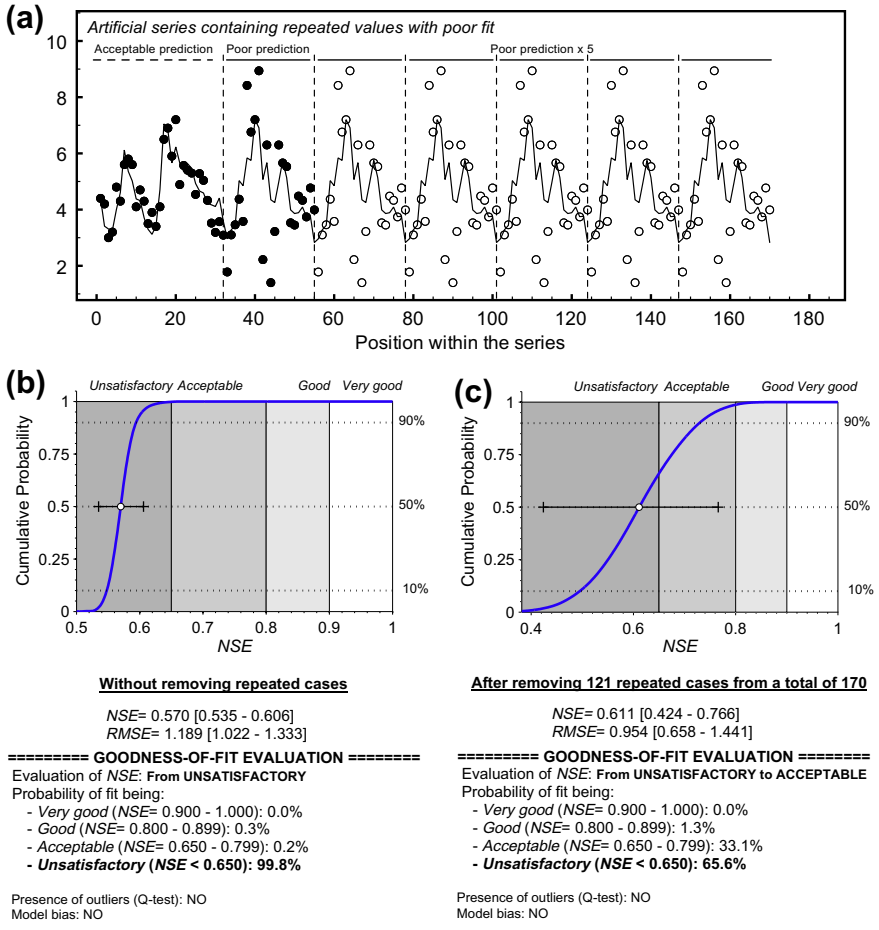

Fig. 9. Effect of repeated cases in the observed and predicted sets on the goodness-of-fit assessment. When repeated values worsen the goodness-of-fit. Outputs were obtained using the computer tool presented in the Appendix.

Finally, the effect of repeated values in the dataset is illustrated with nitrate data from well *W10* (Fig. 4b). The concentration measured in this well (solid symbols in Fig. 8a) was better predicted in the first 32 values than the remaining 23 values. To explore the effect of repeated values, we synthesized a new time series by replicating the first 32 values four times (Fig. 8a), followed by the last 23 values in the original series. The goodness-of-fit evaluation was then conducted by comparing the series containing the repeated values (Fig. 8b) and the original series (Fig. 8c). Note how the repeated values have a positive influence on the *NSE* and on its significance (*NSE* = 0.704 [0.598–0.774] in Fig. 8b vs. *NSE* = 0.611 [0.424–0.766] in Fig. 8c). In addition, the information in Fig. 8b suggests that the fit may be considered significantly valid (*p*-value = 0.089) while when the repeated values are not taken into account (Fig. 8c), the evaluation indicates that the fit should be rejected, (*p*-value = 0.656). This clearly illustrates that the repeated cases in this example lead to an overestimation of the model performance. Similarly, the opposite effect is shown by generating artificially new sets of observations and estimates by adding to the original series the last 23 worst-fitting values five times (Fig. 9a). In this case, taking into account the repeated values in the goodness-of-fit evaluation reduces the model performance, where the probability of the fit being *Unsatisfactory* increases to 99.8% (Fig. 9b vs. c). Notice that results in Fig. 4b do not match those of Figs. 8c and 9c, because data from well *W10* contained repeated values. Although these synthetic examples may be viewed as unrealistic or extreme cases, they clearly illustrate how the goodness-of-fit evaluation could be affected by the presence of repeated values in the observed and computed series. Notice however, that removing repeated values from the data sets would make sense only if these were the result of similar physical processes.

4. Conclusions

Hydrological and other environmental models evaluation is enhanced when conducting a multi-objective analysis. We present here a unified framework for proper interpretation of model performance in a statistically rigorous way and for the evaluation of other effects such as bias, outliers and repeated data. As shown in this work, when the goodness-of-fit evaluation is based on a single indicator like the widely used Nash–Sutcliffe coefficient of efficiency (*NSE*), modelers must bear in mind that model error is not linearly related with the indicator value, and that the value is affected by other factors (outliers, model bias, repeated data). Hypothesis testing of the *NSE* exceeding threshold values is also accomplished based on approximated probability distributions obtained by boostrapping or block blockstrapping in the case of time series data. A comprehensive procedure for evaluating model performance is proposed and tested here that can serve as a useful guidance and less subjective tool for modelers. Through its computer implementation (see Appendix), the unified evaluation procedure is simple and quick and allows the modeler to reduce subjective in the goodness of fit statistics interpretation, to accept or reject the model calibration at a desired significance level. Additionally, the modeler can assess the effect that biased model predictions and the presence of outliers or repeated values in the series might have on the goodness-of-fit evaluation. Although focused on the original NSE, the evaluation procedure is also compatible with modified forms of the coefficient of efficiency (Le Moine, 2008; Legates and McCabe, 1999; Oudin et al., 2006). The proposed framework is not intended to prescribe specific reference values to use in the model evaluation process. Rather the procedure is robust and allows comparison of the model performance across applications, or with other models on the same application. This additional transparency offers the potential to advance current modeling practice by limiting the subjectivity commonly introduced in the evaluation of the numerical values of the goodness-of-fit indicators.

Acknowledgements

This work was supported by Project RTA2009-161 funded by the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Ministerio de Ciencia e Innovación. RMC wishes to acknowledge support of the S1042 USDA CSREES Regional Project.

Appendix A. Description of the computer application for model evaluation (FITEVAL)

The generic procedure presented for the evaluation of the goodness-of-fit of mathematical models has been implemented in MATLAB®. The code, called FITEVAL, is available free of charge as a computer application (MS-Windows® and MacIntosh®) or MATLAB function at http://aritter.webs.ull.es/software.html and at http://abe.ufl.edu/carpena/software/FITEVAL.shtml. The code uses a more general formulation of the coefficient of efficiency (Eq. (6)), thus allowing the user to compute modified versions of this indicator instead of the *NSE*.

$$E_j = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^j}{\sum_{i=1}^{N}|O_i - B_i|^j} \tag{6}$$

where B_i is a benchmark series, which may be a single number (such as the mean of observations), seasonally varying values (such as seasonal means), or predicted values using a function of other variables. For $j = 2$ and $B_i = \overline{O}$, Eq. (6) yields $E_2 = NSE$ (used as default).

FITEVAL requires an ASCII text file located in the same directory as the application containing two paired vectors or columns: the first with the observations (O_i) and the second with the model-calculated values (P_i). If the user wants FITEVAL to compute Eq. (5) using a benchmark series instead of $B_i = \overline{O}$, this should be included as a third column in the ASCII text file.

FITEVAL uses MATLAB built-in functions to perform bootstrapping, according to Efron and Tibshirani (1993), and the bias corrected and accelerated calculation of confidence intervals. Implementation of Politis and Romano (1994) stationary block bootstrap (default method) was accomplished using the function available in Kevin Sheppard's Oxford MFE Toolbox (available at http://www.kevinsheppard.com/wiki/MFE_Toolbox). In addition, the automatic selection of bootstrap block length based on Politis and White (2004) algorithm was implemented using Andrew Patton's code, available at http://public.econ.duke.edu/~ap172/code.html.

The program is flexible to be used with other model efficiency threshold values than those proposed in this work, or for calculating Legates and McCabe (1999) modified form of the coefficient of efficiency (*E1*) instead of *NSE*. For this purpose an optional ASCII text file (named fitevalconfig.txt) could be included in the working directory. This contains seven lines corresponding to: *Acceptable* $NSE_{threshold}$, *Good* $NSE_{threshold}$, *Very good* $NSE_{threshold}$, relative bias threshold value (%), the option for computing *E1*, bootstrap type, figures' font size, and the option for canceling the on-screen display of the graphical output. The corresponding default values are 0.65, 0.80, 0.90, 5, 0, 0, 10, and 0, respectively. Notice that using transformed series in the corresponding ASCII text file, such as $\{\sqrt{O_i}, \sqrt{P_i}\}$, $\{\ln(O_i + \varepsilon), \ln(P_i + \varepsilon)\}$ or $\left\{\frac{1}{O_i+\varepsilon}, \frac{1}{P_i+\varepsilon}\right\}$, the program computes *NSE* (but also *RMSE*) applied on root squared, log and inverse transformed values, respectively (Le Moine, 2008; Oudin et al., 2006).

When executing FITEVAL (from the command line by typing "FITEVAL *filename*" or by clicking on the application icon under MS-Windows®), it performs the goodness-of-fit evaluation producing a graphical output presented on the screen that is also automatically saved into a computer file in different graphical formats ('pdf' as default, and 'eps', 'pdf', 'jpg', 'tiff', or 'png' when providing these arguments at the end of the command line execution). Additionally, the numerical output is stored in an ASCII text file. When executing the application with NOREP as argument in the command line (FITEVAL *filename* NOREP), the program will remove all repeated values from the data set before performing the goodness-of-fit evaluation.

The graphical file and its screen presentation contain all the elements for a complete model goodness-of-fit evaluation: (a) a plot of observed vs. computed values illustrating the match on the 1:1 line; (b) the calculation of *NSE* and *RMSE* and their corresponding confidence intervals of 95%; (c) the qualitative goodness-of-fit interpretation based on the established classes; (d) a verification of the presence of bias or the possible presence of outliers; (e) the plot of the *NSE* cumulative probability function superimposed on the *NSE* class regions; and (f) a plot illustrating the evolution of the observed and computed values. The latter plot helps for visually inspecting the similarity degree of the two series, and detecting which observations are best or worst predicted by the model. Examples of the FITEVAL output are presented in Figs. 4–9.

References

Bárdossy, A., Singh, S.K., 2008. Robust estimation of hydrological model parameters. Hydrol. Earth Syst. Sci. 12, 1273–1283.

Beven, K.J., 2006. A manifesto for the equifinality thesis. J. Hydrol. 320, 18–36.

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., Montanari, A., 2012. Validation of hydrological models: conceptual basis, methodological approaches and a proposal for a code of practice. Phys. Chem. Earth 42–44, 70–76.

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

Carlstein, E., 1986. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann. Statist. 14, 1171–1179.

Clark, M.P., Kavetski, D., Fenicia, F., 2011a. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour. Res. 47, W09301. http://dx.doi.org/10.1029/2010WR009827.

Clark, M.P., McMillan, H.K., Collins, D.B.G., Kavetski, D., Woods, R.A., 2011b. Hydrological field data from a modeller's perspective. Part 2: process-based evaluation of model hypotheses. Hydrol. Process. 25, 523–543.

Clarke, R.T., 2008. A critique of present procedures used to compare performance of rainfall–runoff models. J. Hydrol. 352, 379–387.

Criss, R.E., Winston, W.E., 2008. Do Nash values have value? Discussion and alternate proposals. Hydrol. Process. 22, 2723–2725.

Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web–based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Model. Software 22 (7), 1034–1052.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Stat. Sci. 11, 189–228.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.

Efron, B., 1981a. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika 68, 589–599.

Efron, B., 1981b. Nonparametric standard errors and confidence intervals. Can. J. Statist 9, 139–158.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Ehret, U., Zehe, E., 2011. Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrol. Earth Syst. Sci. 15, 877–896.

Ewen, J., 2011. Hydrograph matching method for measuring model performance. J. Hydrol. 408, 178–187.

Garrick, M., Cunnane, C., Nash, J.E., 1978. A criterion of efficiency for rainfall-runoff models. J. Hydrol. 36, 375–381.

Guinot, V., Cappelaere, B., Delenne, C., Ruelland, D., 2011. Towards improved criteria for hydrological model calibration: theoretical analysis of distance- and weak form-based functions. J. Hydrol. 401, 1–13.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrological models: multiple and noncommensurable measures of information. Water Resour. Res. 34, 751–763.

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. J. Hydrol. 377, 80–91.

Gupta, H.V., Kling, H., 2011. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. Water Resour. Res. 47, W10601. http://dx.doi.org/10.1029/2011WR010962.

Harmel, R.D., Smith, P.K., Migliaccio, K.W., 2010. Modifying goodness-of-fit indicators to incorporate both measurement and model uncertainty in model calibration and validation. Trans. ASABE 53, 55–63.

Houghton-Carr, H.A., 1999. Assessment criteria for simple conceptual daily rainfall–runoff models. J. Hydrol. Sci. 44 (2), 237–261.

Jain, S.K., Sudheer, K.P., 2008. Fitting of hydrologic models: a close look at the Nash–Sutcliffe Index. J. Hydrol. Eng. 13, 981–986.

James, L.D., Burgess, S.J., 1982. Selection, calibration and testing of hydrologic models. In: Haan, C.T., Johnson, H.P., Brakensiek, D.L. (Eds.), Hydrologic Modeling of Small Watersheds. ASAE, St. Joseph, Mich, pp. 437–472.

Kapetanios, G., Papailias, F., 2011. Block Bootstrap and Long Memory. Working Paper 679, Queen Mary University of London.

Kavetski, D., Clark, M.P., 2011. Numerical troubles in conceptual hydrology: approximations, absurdities and impact on hypothesis-testing. Hydrol. Process. 25, 661–670.

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.

Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. Ann. Statist. 17, 1217–1241.

Legates, D.R., Davis, R.E., 1997. The continuing search for an anthropogenic climate change signal: limitations of correlation-based approaches. Geophys. Res. Lett. 24, 2319–2322.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241. http://dx.doi.org/10.1029/1998WR900018.

Legates, D.R., McCabe, G.J., 2012. Short communication a refined index of model performance. A rejoinder. Int. J. Climatol.. http://dx.doi.org/10.1002/joc.3487.

Le Moine, N., 2008. Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performance et du réalisme des modéles pluie–débit? PhD Thesis, Université Pierre et Marie Curie, Antony, 324 pp.

Liu, R.Y., Singh, K., 1992. Moving blocks jackknife and bootstrap capture weak dependence. In: Le Page, R., Billard, L. (Eds.), Exploring the Limits of Bootstrap. Wiley, New York, pp. 225–248.

Loague, K., Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. J. Contaminant Hydrol. 7, 51–73.

Martinec, J., Rango, A., 1989. Merits of statistical criteria for the performance of hydrological models. J. Am. Water Resour. Assoc. 25 (2), 421–432.

McCuen, R.H., Snyder, W.M., 1975. A proposed index for comparing hydrographs. Water Resour. Res. 11 (6), 1021–1024.

McCuen, R.H., 2003. Modeling Hydrologic Change: Statistical Methods. CRC Press, Boca Raton, FL.

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency Index. J. Hydrol. Eng. 11, 597–602.

McMillan, H.K., Clark, M.P., Bowden, W.B., Duncan, M., Woods, R.A., 2011. Hydrological field data from a modeler's perspective. Part 1: diagnostic tests for model structure. Hydrol. Process. 25, 511–522.

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. Trans. ASABE 50, 885–900.

Moussa, R., 2010. When monstrosity can be beautiful while normality can be ugly: assessing the performance of event–based flood models. J. Hydrol. Sci. 55 (6), 1074–1084.

Muñoz-Carpena, R., Ritter, A., Li, Y.C., 2005. Dynamic factor analysis of groundwater quality trends in an agricultural area adjacent to Everglades National Park. J. Contam. Hydrol. 80, 49–70.

Murphy, A., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Weather Rev. 116, 2417–2424.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, part I: a discussion of principles. J. Hydrol. 10, 282–290. http://dx.doi.org/10.1016/0022-1694(70)90255-6.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., 2006. Dynamic averaging of rainfall–runoff model simulations from complementary model parameterizations. Water Resour. Res. 42 (7), W07410.

Patton, A., Politis, D.N., White, H., 2009. Correction to "automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. Econom. Rev. 28, 372–375.

Politis, D.N., Romano, J.P., 1992. A circular block-resampling procedure for stationary data. In: LePage, R., Billard, L. (Eds.), Exploring the Limits of Bootstrap. John Wiley, New York, pp. 263–270.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Amer. Statist. Assoc. 89, 1303–1313.

Politis, D.N., White, H., 2004. Automatic block-length selection for the dependent bootstrap. Econom. Rev. 23, 53–70.

Pushpalatha, R., Perrin, C., Le Moine, N., Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. J. Hydrol. 420–421, 171–182.

Reusser, D.E., Blume, T., Schaefli, B., Zehe, E., 2009. Analysing the temporal dynamics of model performance for hydrological models. Hydrol. Earth Syst. Sci. 13 (7), 999–1018.

Ritter, A., Muñoz-Carpena, R., Bosch, D.D., Schaffer, B., Potter, T.L., 2007. Agricultural land use and hydrology affect variability of shallow groundwater nitrate concentration in South Florida. Hydrol. Process. 21, 2464–2473.

Rorabacher, D.B., 1991. Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level. Anal. Chem. 63, 139–146.

Santhi, C., Arnold, J.G., Williams, J.R., Dugas, W.A., Srinivasan, R., Hauck, L.M., 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources. J. Am. Water Resour. Assoc. 37, 1169–1188.

Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21 (15), 2075–2080.

Seibert, J., 2001. On the need for benchmarks in hydrological modelling. Hydrol. Process. 15 (6), 1063–1064.

Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration. Water Resour. Res. 38 (11), 1241.

Singh, S.K., Bárdossy, A., 2012. Calibration of hydrological models on hydrologically unusual events. Adv. Water Resour. 38, 81–91.

Van Liew, M.W., Arnold, J.G., Garbrecht, J.D., 2003. Hydrologic simulation on agricultural watersheds: choosing between two models. Trans. ASAE 46, 1539–1551.

Wagener, T., 2003. Evaluation of catchment models. Hydrol. Process. 17, 3375–3378.

Willems, P., 2009. A time series tool to support the multi-criteria performance evaluation of rainfall–runoff models. Environ. Modell. Software 24, 311–321.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90, 8995–9005.

Young, P.C., Ratto, M., 2009. A unified approach to environmental systems modeling. Stoch. Env. Res. Risk Assess. 23 (7), 1037–1057.

Zuur, A.F., Fryer, R.J., Jolliffe, I.T., Dekker, R., Beukema, J.J., 2003. Estimating common trends in multivariate time series using dynamic factor analysis. Environmetrics 14, 665–685.