Hydrol. Earth Syst. Sci., 16, 3049–3060, 2012
www.hydrol-earth-syst-sci.net/16/3049/2012/
doi:10.5194/hess-16-3049-2012

Hydrology and Earth System Sciences

Ideal point error for model assessment in data-driven river flow forecasting

C. W. Dawson[1], **N. J. Mount**[2], **R. J. Abrahart**[2], **and A. Y. Shamseldin**[3]

[1]Department of Computer Science, Loughborough University, Loughborough, UK
[2]School of Geography, University of Nottingham, Nottingham, UK
[3]Department of Civil and Environmental Engineering, University of Auckland, Auckland, New Zealand

Correspondence to: A. Y. Shamseldin (a.shamseldin@auckland.ac.nz)

Received: 13 November 2011 – Published in Hydrol. Earth Syst. Sci. Discuss.: 6 February 2012
Revised: 13 July 2012 – Accepted: 13 July 2012 – Published: 29 August 2012

Abstract. When analysing the performance of hydrological models in river forecasting, researchers use a number of diverse statistics. Although some statistics appear to be used more regularly in such analyses than others, there is a distinct lack of consistency in evaluation, making studies undertaken by different authors or performed at different locations difficult to compare in a meaningful manner. Moreover, even within individual reported case studies, substantial contradictions are found to occur between one measure of performance and another. In this paper we examine the ideal point error (IPE) metric – a recently introduced measure of model performance that integrates a number of recognised metrics in a logical way. Having a single, integrated measure of performance is appealing as it should permit more straightforward model inter-comparisons. However, this is reliant on a transferrable standardisation of the individual metrics that are combined to form the IPE. This paper examines one potential option for standardisation: the use of naive model benchmarking.

1 Introduction

Evaluation metrics that provide a quantitative comparison of the fit between an observed hydrological record and a model's prediction are a central component of validation. Their correct use and interpretation is fundamental for the justification of a model and the evaluation of its performance in a given hydrological application. It is of no surprise, therefore, that evaluation metrics have received considerable attention in the last decade with respect to their usefulness and applicability in different hydrological modelling contexts (e.g. Beran, 1999; Seibert, 2001; Criss and Winston, 2008; Jain and Sudheer, 2008). Schaefli and Gupta (2007) stressed that hydrological model evaluation metrics were important, not only as an integral part of model development and calibration processes, but also as a means of communicating results to scientists, stakeholders and other end-users. This recognition has led to renewed interest in model evaluation metrics by hydrologists, in which the basic approach has been to undertake detailed exploration of individual metrics and their interpretation in different theoretical or real-world scenarios. Indeed, six studies afforded to the Nash-Sutcliffe efficiency index (NSE: Nash and Sutcliffe, 1970) in the last five years are a prime example of the desire by hydrologists to better understand the evaluation metrics that they are using (Schaefli and Gupta, 2007; Criss and Winston, 2008; Jain and Sudheer, 2008; Gupta et al., 2009; Ruesser et al., 2009; Moussa, 2010).

In data-driven modelling, the importance of evaluation metrics is particularly acute since there are few other mechanisms available by which the performance of a data-driven model (DDM) can be assessed. Moreover, the implicit black-box characteristic of many DDMs prevents the examination and verification of the physical rationality of the modelling mechanisms (Minns and Hall, 1996; Babovic, 2005; Abrahart et al., 2012a,b). Consequently, the majority of DDM studies in hydrology identify the best or preferred model solely on the basis of superior metric score(s) achieved, even though this approach has been argued to be overly simplistic (Abrahart et al., 2011). In addition, the fact that there is a wide range of potential sources of error in hydrological models that impact differently on different performance metrics (Criss and Winston, 2008; Willems, 2012) makes the choice of which metric to choose complex. It is, therefore, vital that researchers provide adequate clarification of what the specific values of different metrics really mean in the context of the errors that may be present in their models. It is also vital that assessments are not made on the basis of an individual evaluation metric score as it may be unduly influenced by a specific error component that is present in the model output (American Society of Civil Engineers, 1993).

In recognition of this, it has become standard practice for data-driven modellers to assess the performance of their models against a large number of different metrics with a host of different combinations potentially being applied to a particular solution (Elshorbagy et al., 2000). However, there is little consistency in how different metrics are adopted from one study to another (Legates and McCabe, 1999), a fact that may be down to the provision of different default metrics in modelling software packages (Chiew and McMahon, 1993). As Hall (2001) pointed out: "*Ideally, the modeller would wish to express the goodness-of-fit of the model to the data in terms of a single index or objective function.*" Although researchers have acknowledged the importance of multi-criteria performance analysis (for example, Masmoudi and Habaieb, 1993; Weglarczyk, 1998; Willems, 2009), developments in the integration of multiple error measures into a single measure of hydrological model performance have only recently received attention, and their application in data-driven studies remains rare.

Combining multiple evaluation metric scores into a single function raises an important technical question about how very different scales and value ranges associated with individual metrics can be standardised. Gupta et al. (2009) proposed the Kling-Gupta efficiency index (KGE), which delivers a measure of Euclidean distance from an ideal point in "scaled space". Their approach standardises the component metrics on the basis of each model's deviation from the mean and standard deviation of the observed data series. This results in a flexible integrated metric that can be computed using either un-weighted or re-scalable equations. It also offers the potential to fine-tune the metric so that it responds more or less strongly to different error types. However, their use of statistical parameters as the basis for a standardisation means that consistent statistical distributions of error are expected if models are to be compared using KGE, and this assumption may not always be met. More recently, Elshorbagy et al. (2010a,b) proposed the ideal point error (IPE). This metric builds upon the idea of Gupta et al. (2009) of quantifying the distance from an ideal point. However, it uses the deviation of a model's multiple goodness-of-fit metrics from their "perfect" scores as a standard (in which a perfect match between the modelled and observed series results in a perfect score of zero), rather than statistical measures of deviation. This arguably results in a more flexible evaluation tool, which can integrate a wider range of metrics, and makes no assumptions about the error distributions of the different models being compared. However, difficulties remain when comparing IPE scores for different models, particularly if the relative performance of models that have been developed to predict different observed series is of interest. Each metric in IPE, and hence the final IPE score itself, is unique to a particular data set, such that a metric standardisation process based on the observed series will not be consistent across different modelling applications and contexts.

An alternative is to use a simple model, rather than the observed data set, as the basis for metric standardisation. In this way, the metric scores for a given model are compared to those obtained by a simple, standard, baseline model. Each model can then be assessed with respect to its relative performance gain over the baseline model. This idea is not new and is the basic concept underpinning NSE, in which model performance is assessed relative to a highly simplistic model that represents the mean of the observed record. Metrics that are standardised in this way do not enable direct comparison of values from models developed on independent data in any absolute terms. However, they do have the important advantage of enabling a transferrable comparison of each model's relative performance gain over a baseline model type (e.g. Seibert, 2001; Moussa, 2010) and this will be consistent from study to study. Despite these advantages, standardisation of metrics in this way remains rare and few studies have examined the impact of model benchmarking with respect to the differential performance of a specific hydrological modelling evaluation metric.

Given the potential benefits of metric standardisation using simple models, it was perhaps surprising that this approach was not considered in the original papers of Elshorbagy et al. (2010a,b) as a method for standardising the metrics in IPE. Indeed, only limited discussion and evaluation of selection and integration procedures were provided in the source articles. As a consequence, two key questions remain unaddressed:

1. How is IPE output impacted by the particular composition and distribution of errors in the suite of models under test?

2. What is the impact of standardising IPE to a baseline model, rather than to the observed data?

This paper responds to the current interest in better understanding hydrological metrics by undertaking a detailed assessment of potential strengths and weaknesses associated with IPE in the context of a simple river flow forecasting application. The significance of such research is twofold: in performing an examination of potential benefits and limitations surrounding the use of a new evaluation metric, and in exploring the adoption of simple models as standardisation baselines. The above questions are studied by

1. revisiting the method of Elshorbagy et al. (2010a,b) and developing general rules for error metric inclusion in

IPE, together with a simple variant equation that avoids several numerical problems encountered when applying the original;

2. assessing the output consistency of the original IPE equation, the variant outlined above and the variant proposed by Dominguez et al. (2011) in which metric orthogonality is enforced;

3. examining the impact of standardising IPE to a naive, autoregressive baseline model.

Individual statistics in this paper were calculated using HydroTest (www.hydrotest.org.uk): a standardised, open access website that performs the required numerical calculations (Dawson et al., 2007, 2010). Equations and sources for the different metrics are provided in its related papers and web pages. The following abbreviations for each computed metric will be employed in the remainder of this paper: root mean squared error (RMSE), mean absolute relative error (MARE), mean error (ME), correlation coefficient (R), R-squared (RSqr), persistence index (PI), percentage error in peak (PEP) and NSE (defined earlier; labelled CE in HydroTest, but now rebadged as NSE, in order to avoid any possible conflict and/or confusion arising from the fact that our discipline now possesses two alternative measures of hydrological modelling efficiency NSE and KGE).

2 Ideal point error

2.1 Metric standardisation

IPE is a dimensionless composite index that measures model performance with respect to an ideal point in an n-dimensional space (where n is the number of model performance evaluation metrics employed). It standardises a set of model performance evaluation statistics to an ideal point lying at [0, 0, 0, ..., 0]. The worst case is at [1, 1, 1, ..., 1]. The overall performance of a model in terms of IPE is measured as the Euclidian distance from that ideal point (i.e. smaller is better). If IPE is applied to a group of model outputs computed on the same data set, an IPE value of unity corresponds to the worst performing model; an IPE value of zero corresponds to a perfect (ideal) model. Elshorbagy et al. (2010a,b) published an original IPE index (here termed $\mathrm{IPE_A}$) that integrated four popular metrics (in which they referred to ME as mean bias, MB):

$$\mathrm{IPE_A} = \left[0.25\left(\left(\frac{\mathrm{RMSE}_i}{\max(\mathrm{RMSE})}\right)^2 + \left(\frac{\mathrm{MARE}_i}{\max(\mathrm{MARE})}\right)^2 + \left(\frac{\mathrm{ME}_i}{\max|\mathrm{ME}|}\right)^2 + \left(\frac{R_i - 1}{1/\max(R)}\right)^2\right)\right]^{1/2} \quad (1)$$

where, for model i, max (x) is the maximum value of the statistic x among the group of models under test and is used as a standardisation factor of model performance for each individual assessment metric. The four selected error statistics, along with a visual comparison performed between observed and predicted values, were considered to be sufficient to reveal any significant differences occuring amongst the various modelling approaches being compared with regard to their prediction accuracy.

Table 1. Fivefold classification of potential components used in IPE.

Category	Examples	Best	Worst	IPE component
S1	RMSE, MARE	0	$+\infty$	$\frac{\mathrm{S1}}{\max(\mathrm{S1})}$
S2	RSqr	1	0	$\frac{\mathrm{S2}-1}{\min(\mathrm{S2})-1}$
S3	R	1	-1	$\frac{\mathrm{S3}-1}{\min(\mathrm{S3})-1}$
S4	PEP, ME	0	$\pm\infty$	$\frac{\mathrm{S4}}{\max\lvert\mathrm{S4}\rvert}$
S5	NSE, PI	1	$-\infty$	$\frac{\mathrm{S5}-1}{\min(\mathrm{S5})-1}$

Detailed inspection of the original IPE equation (Eq. 1) reveals that it is inconsistent in the method of standardisation which is carried out for each component. The first three metrics are standardised with respect to the worst performing model, while the last metric (R) is standardised with respect to the best performing model. It should also be noted that the reported standardisation of the correlation coefficient (R) presented in Eq. (1) is not designed to deal with negative scores which could deliver components that exceed the maximum upper limit for a perfect score (i.e. > 1).

One of the key advantages of IPE is the flexibility with which it can accommodate a wide range of different error metrics. However, care must be taken over the exact manner in which specific metrics are integrated. Table 1 summarises how certain classes of error measure should be standardised for integration into an IPE. These classes, referred to as S1–S5, are based on the range of potential outputs for a particular metric (best and worst).

Equation (2) represents an improved variant (here termed $\mathrm{IPE_B}$) of the original equation, which includes a more generalised and robust procedure for standardising R that can accommodate its full range [−1, +1]. $\mathrm{IPE_B}$ is also consistent in the way that standardisation is performed with respect to the worst performing model. Thus, it eliminates the standardisation inconsistency of Eq. (1). This modification can result in a significant difference arising between the output of $\mathrm{IPE_A}$ and $\mathrm{IPE_B}$, particularly for situations containing moderate or low correlation coefficient values. Indeed, as the results presented later show, correlation coefficient scores as high as 0.91 can still result in quite different scores for $\mathrm{IPE_A}$ and $\mathrm{IPE_B}$.

$$\mathrm{IPE_B} = \left[0.25\left(\left(\frac{\mathrm{RMSE}_i}{\max(\mathrm{RMSE})}\right)^2 + \left(\frac{\mathrm{MARE}_i}{\max(\mathrm{MARE})}\right)^2 + \left(\frac{\mathrm{ME}_i}{\max|\mathrm{ME}|}\right)^2 + \left(\frac{R_i - 1}{\min(R) - 1}\right)^2\right)\right]^{1/2} \qquad (2)$$

2.2 Equifinal models

A further potential difficulty with IPE arises in the case of equifinal models, which are known to be a problem in the field of hydrology (Beven, 1993, 1996, 2001). Equifinal models will result in IPE values close to unity for each model, indicating (possibly incorrectly) that all models are poor because of the manner in which IPE is derived relative to the worst performing model in the suite of models under evaluation. However, if, as suggested later in this paper, IPE is based on a common benchmark (such as a naive model), then all models are compared to that benchmark rather than against one another and the problem is alleviated. In addition, if IPE still produces similar values for different models, this would simply be highlighting the equifinal nature of the models under test. In this situation, detailed inspection of the corresponding hydrograph might possibly tease out subtle differences arising between individual solutions.

2.3 Metric orthogonality

In applying any integrated evaluation metric, the question of which set of metrics to use is central. One approach to answering this question is to consider the extent to which the metrics overlap one another with respect to their discriminatory power. Dominguez et al. (2011) published a modified IPE index (here termed $\mathrm{IPE_C}$) which integrated five popular metrics ordered according to their power of appraisal. It was strongly argued in their paper that the individual statistics which are selected for inclusion in such procedures should be orthogonal (i.e. uncorrelated), as well as comprehensive, to avoid potential issues of information redundancy (i.e. loss of discriminatory power) and/or double-counting (i.e. multiple accumulated measures that assess identical factors). Thus, following detailed analysis of numerous potential candidates, only two of the four original $\mathrm{IPE_A}$ metrics were retained in their modified equation (RMSE and ME) and R was replaced by RSqr:

$$\mathrm{IPE_C} = \left[0.2\left(\left(\frac{\mathrm{RMSE}_i}{\max(\mathrm{RMSE})}\right)^2 + \left(\frac{\mathrm{RSqr}_i - 1}{\min(\mathrm{RSqr}) - 1}\right)^2 + \left(\frac{\mathrm{ME}_i}{\max|\mathrm{ME}|}\right)^2 + \left(\frac{\mathrm{PI}_i - 1}{\min(\mathrm{PI}) - 1}\right)^2 + \left(\frac{\mathrm{PEP}_i}{\max|\mathrm{PEP}|}\right)^2\right)\right]^{1/2} \qquad (3)$$

$\mathrm{IPE_C}$ was derived from an examination of 22 different statistical metrics for each of 60 models. Principal component analysis (PCA) was used to derive surrogate measures of performance that encapsulated the information contained in all 22 statistical metrics. The first five components provided 91 % of the information content of all 22 metrics. These orthogonal components were then examined to determine which metrics could best represent them. Subsequent analysis led to the five metrics used in Eq. (3): such selections, it should be noted, being dependent on the data set involved.

The use of a comprehensive PCA approach to analyse orthogonality is not always going to be feasible, particularly if only a few models and metrics are being compared. In such circumstances, a basic correlation analysis should be sufficient to detect redundant metrics which, if not removed, would bias IPE output (i.e. identification and removal of highly correlated metrics is recommended). Performing such an analysis would seem to be a prudent early step in all applications of IPE, and one which can quickly identify the best number and mix of metrics to be included. We, therefore, perform just such an analysis in our evaluation of IPE later in this paper. In cases where only a few individual models are being compared, there may be an insufficient number of models present to perform a meaningful analysis that could identify redundant IPE component metrics. Instead, the selection of metrics will need to be made on the basis of results collated from previously published hydrological modelling studies performed for similar forecasting scenarios conducted in a similar environmental setting.

3 Numerical experiments

In this paper, we explore IPE and its variants in the context of a simple river flow forecasting application, a topic that has been a focus of data-driven hydrological modelling studies over the last two decades, e.g. in neural network modelling (Maier et al., 2010). We use artificial errors here to engineer a set of "models" from an observed data series, with each model having one characteristic error type associated with it. In this way, we provide a theoretical insight into how different variants of IPE behave when faced with different characteristic errors in the fit between observed and predicted series. Our approach is supported by several examples of studies in which artificially engineered "synthetic errors" are computed for a simple hydrological data set, and used as the basis for examining the variation in evaluation metric scores under different theoretical scenarios (e.g. Krause et al., 2005; Cloke and Pappenberger, 2008). Although a simplification of the real-world, this approach is justified by the extent to which reported results can be interpreted in both simple and fundamental terms, thereby offering a degree of general transferability for our findings. This key outcome and such clarity would be impossible to accomplish from analyses involving multifaceted, real-world case studies possessing complex, compound errors.

The observed record used in this study was first adopted as an instrument for performing error testing operations in Dawson and Wilby (2001). It relates to six-hourly discharge recorded in $m^3 s^{-1} \times 10^2$ at the site of the Three Gorges Dam, on the Yangtze River in China. The data comprise 160 observed records for the period 4 July 1992 to 13 August 1992 and are depicted in Fig. 1. This data set can be downloaded from the HydroTest website (Dawson et al., 2007, 2010). Further particulars on the origins of the data set can be found in Dawson et al. (2002).

The IPE variants specified in Eqs. (1)–(3) are evaluated and benchmarked in this paper using 12 simple data series, which are compared against the observed record. The first four data series are generated from simple models of the observed record: two naive time-shift models (as used by Hall, 2001); and two simple linear regression models. The other eight data series were constructed by introducing different types of error into the observed record. The first four of these are based on the ones used by Hall (2001) in his evaluation of popular goodness-of-fit indices. The second four involve the use of random numbers sampled from a normal distribution.

Two versions of each data series are included, representing large and small deviations from the observed record. The formulae used to calculate the modified records are given in Eqs. (4)–(9) below (in which $\hat{Q}_i$ is the estimated discharge):

1. Two naive time-shift models that forecast observed discharge. This series can be expressed as:

   $$\hat{Q}_i = Q_{i-n} \tag{4}$$

   in which n is the lag-time. In this case two lag times are used: a lag of one ($n = 1$) representing a 6 h, 1 step-ahead naive forecast; and a lag of four ($n = 4$) representing a 24 h, 4 step-ahead naive forecast. These models are referred to as Naive ($t + 1$) and Naive ($t + 4$).

2. Two simple linear regression models that use antecedent flow as a predictor for delivering $t + 1$ step-ahead and $t + 4$ step-ahead forecasts of observed discharge (and which are consistent with our naive modelling solutions Naive ($t + 1$) and Naive ($t + 4$)). This series can be expressed as:

   $$\hat{Q}_i = r_n Q_{i-n} + k_n \tag{5}$$

   where r_n is the regression coefficient for time lag n, and k_n is the intercept. These are referred to as Regression ($t + 1$) and Regression ($t + 4$). For $n = 1$, $r_1 = 0.999$ and $k_1 = -0.332$. For $n = 4$, $r_4 = 0.927$ and $k_4 = 18.324$. Note that r_1 is close to unity.

3. Synthetic series containing scaled errors that are proportional to the magnitude of the observed flow. This series can be expressed as:

   $$\hat{Q}_i = c\, Q_i \tag{6}$$

   where c is a constant. Two values of c are adopted in this paper to assess the effects of varying degrees of error: 1.25 and 1.5. The latter is the upper value applied by Hall (2001). The former represents half that applied error. These series are referred to as Scaled (low) for $c = 1.25$ and Scaled (high) for $c = 1.5$.

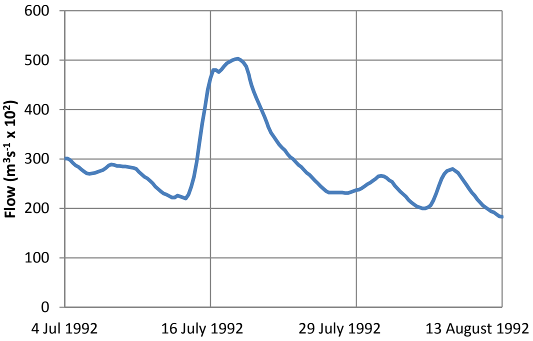

Fig. 1. Hydrograph of observed flow for Three Gorges Dam, Yangtze River, China.

4. Synthetic series containing bias errors in which the observed discharge has been incremented by a constant amount (b) according to the following equation and as such equates to a vertical displacement of the original record. This series can be expressed as:

   $$\hat{Q}_i = Q_i + b. \tag{7}$$

   In order to show how an IPE can differentiate between similar models, b is set to values such that the RMSE of the bias errors is the same as the RMSE of the two scaled errors introduced in Eq. (6) above. In the case of Scaled (low), $b = 74.3$. In the case of Scaled (high), $b = 148.6$. These series are referred to as Bias (low) and Bias (high) respectively.

5. Synthetic series in which random noise has been added to the observed record. This series can be expressed as:

   $$\hat{Q}_i = Q_i + N \tag{8}$$

   in which N is a random value from a normal distribution with a mean of zero and either one or other of two permitted standard deviations. In one case, the standard deviation adopted is one quarter of the standard deviation of the observed record ($N = 20.45$). In the other case, the standard deviation adopted is half that of the standard deviation of the observed record ($N = 40.90$). These values were chosen as they represent a reasonable distribution of noise without generating negative flow values. These series are referred to as Noise (low) and Noise (high) respectively.

6. Synthetic series in which the random noise added to the observed record in Eq. (8) above has been scaled by the

Table 2. Individual statistics for experimental data series ("best" result in bold, "worst" result in italic per metric).

Error model	ME	RMSE	PEP	MARE	RSqr	PI	R
Naive ($t+1$)	0.70	9.24	**0.00**	0.02	0.99	**0.00**	0.99
Naive ($t+4$)	2.85	35.01	**0.00**	0.08	0.83	−13.29	0.91
Regression ($t+1$)	**0.08**	**9.21**	−0.17	**0.02**	0.99	0.01	0.99
Regression ($t+4$)	0.13	34.39	−3.66	0.08	0.83	−12.79	0.91
Scaled (low)	71.38	74.30	25.00	0.25	**1.00**	−63.38	**1.00**
Scaled (high)	142.75	*148.60*	50.00	0.50	**1.00**	*−256.50*	**1.00**
Bias (low)	74.30	74.30	14.77	0.28	**1.00**	−63.38	**1.00**
Bias (high)	*148.60*	*148.60*	29.54	*0.56*	**1.00**	*−256.50*	**1.00**
Noise (low)	−0.59	20.20	6.46	0.06	0.94	−3.76	0.97
Noise (high)	2.07	39.92	5.69	0.12	0.80	−17.58	0.90
Scaled Noise (low)	−3.79	24.86	18.03	0.06	0.91	−6.21	0.96
Scaled Noise (high)	−0.30	48.09	*29.98*	0.11	*0.77*	−25.97	*0.88*

square of the observed record. This leads to proportionally larger errors at high flows and smaller errors at low flows. This series can be expressed as:

$$\hat{Q}_i = Q_i + N\,Q_i^2/z \qquad (9)$$

in which z is a value chosen to ensure scaled errors do not lead to negative flows. In this case, setting z to the square of the mean of the observed record ($z = 285.82$) leads to acceptable results. The two series are referred to as Scaled Noise (low) and Scaled Noise (high) coinciding with the amount of random noise added in Eq. (8).

The relationship between each data series and observed flow is depicted in Fig. 2. The plots show similar performance of the Naive and Regression models with the two models involving one-step-ahead prediction demonstrating low errors across the range of the observed record. The scaled error series show, not surprisingly, a linear increase in error as observed flow increases, while the bias error series show consistent error across the same range. The two noise series (Noise (low) and Noise (high)) show a reasonably even spread of error across the range of the observed record, while scaled noise displays heteroscedastic error in both cases (Scaled Noise (low) and Scaled Noise (high)).

4 Interpretation of error statistics

4.1 Error statistics of the data series

HydroTest statistics for each data series and all relevant evaluation metrics are provided in Table 2. The analysis reveals no overall "winner" (or "loser") in the sense of one data series possessing a superior (or inferior) result for all seven metrics, providing sound grounds for the application of an IPE. For example, Bias (high) returns R and RSqr scores of one (the maximum score) but is identified as the poorest model according to four other statistics (ME, RMSE, MARE and PI). Similarly, Scaled (high) has unity scores for RSqr and R but also has the worst score for RMSE and PI. Conversely, although Naive ($t+1$) possesses the two best scores for PEP and PI, it does not come out on top according to other measures.

Several other points of interest can be identified. First, note that both Naive models return PEP values of zero (i.e. the best permitted score for this metric). This is because both are generated directly as a time-shift of the observed data and consequently have the same peak value as the observed record. This brings into question the use of individual error measures that can return good or perfect results for very simplistic models, and will often also create a divide by zero problem if used as benchmarks in an IPE. Second, although most error statistics return similar scores for the Naive ($t+1$) and Regression ($t+1$) models, there is a notable difference in the ME score for these two models (0.7 and 0.08 respectively). Clearly, bias is reduced as a result of the $k_1 = -0.332$ intercept since all other factors are more or less identical. Moreover, because this measure is calculated using signed differences between the observed and modelled record, there is also a danger that, even for a poor model, substantial differences will cancel one another out leading to good scores. Third, despite returning perfect scores for RSqr and R, the Scaled and Bias series return very poor PI, ME and RMSE scores compared with the other data series. RSqr and R are not good at identifying scaled and bias errors when evaluating models.

The results confirm a long-standing argument in hydrological modelling that individual error statistics cannot be relied upon to provide an objective measure of model performance. This analysis also highlights the dangers of using individual measures that may provide results that are contradictory to what is actually being measured. It is only when error statistics are compared or combined that an overall picture of model performance emerges.

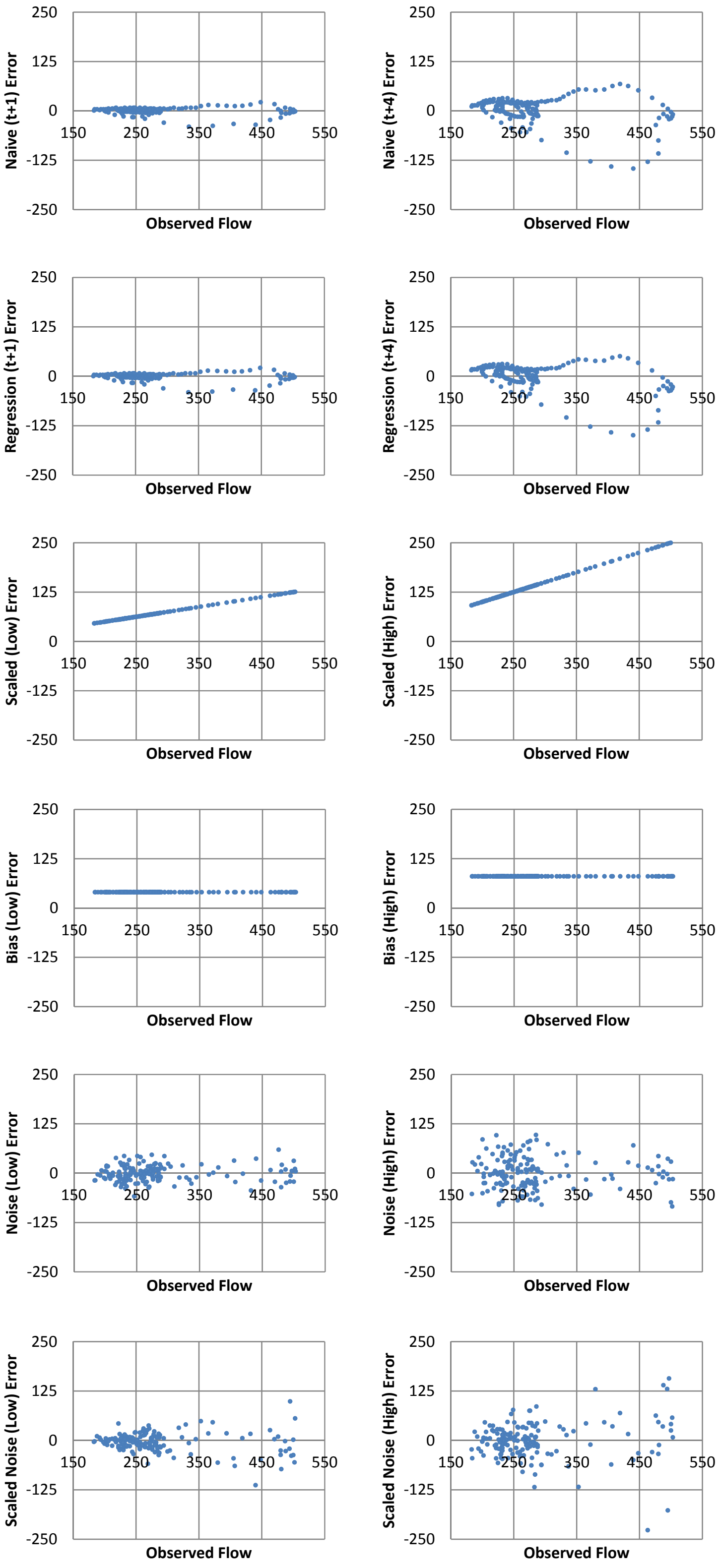

Fig. 2. Error plots for experimental data series (measurements in m^3 s^{-1} $\times$ 10^2).

Table 3. Correlation matrix of individual statistics for experimental data series.

	ME	RMSE	PEP	MARE	RSqr	PI	*R*
ME	1.00						
RMSE	0.97	1.00					
PEP	0.76	0.82	1.00				
MARE	0.98	0.99	0.78	1.00			
RSqr	0.59	0.37	0.27	0.45	1.00		
PI	−0.96	−0.97	−0.79	−0.97	−0.45	1.00	
R	0.60	0.39	0.29	0.47	1.00	−0.46	1.00

4.2 Identification and removal of non-orthogonal metrics

As noted earlier, provided sufficient data are available, it is possible to undertake a cross-correlation analysis between the error metrics under consideration for inclusion in an IPE in order to identify potential metric redundancy. Table 3 provides just such an analysis based on the 12 experimental data series used in this study. These results would tend to indicate some redundancy between ME, RMSE, MARE and PI. However, in this particular case study, the data series have been artificially generated and, as a consequence, perhaps a greater than usual number of series are found to deliver near identical results according to many of the metrics adopted. For example, six of the twelve return almost identical RSqr values, and the bias errors were derived in such a way as to have the same RMSE scores as the scaled errors. If these data represented genuine models in a hydrological study, there would be some argument for removing ME and PI from $\mathrm{IPE_C}$ as they are closely related to RMSE, which would be retained. Conversely, preserving ME and PI may lead to additional emphasis being placed on metrics of this type, perhaps swamping the contribution of other retained metrics such as RSqr and PEP in our study.

In this case, because $\mathrm{IPE_C}$ is based on a comprehensive study of 60 models, we will retain all the components for further analysis. A study of redundancies amongst IPE components is an important issue and should be the subject of further research. Without such an analysis, it is reasonable to accept an IPE such as $\mathrm{IPE_C}$, which is based on a sound hydrological analysis.

5 Evaluating IPE variants

The integrated IPE (A–C) scores for each data series are compared and contrasted in Table 4. The effects of switching from $\mathrm{IPE_A}$ to $\mathrm{IPE_B}$, given varying strengths of correlation coefficient, can be observed. For example, for those data series returning correlation coefficient scores of 1 (Scaled (low), Scaled (high), Bias (low), Bias (high)), there is no change in the scores of $\mathrm{IPE_A}$ and $\mathrm{IPE_B}$. $\mathrm{IPE_A}$ and $\mathrm{IPE_B}$ scores are also the same for the Naive $(t+1)$ and Regression $(t+1)$ models, which both return correlation coefficient scores of 0.99. However, in the case of the Naive $(t+4)$ and Regression $(t+4)$ models, both have correlation coefficient scores of 0.91 and the switch has led to much higher IPE scores: $\mathrm{IPE_A}$ is 0.15 and 0.14 respectively; $\mathrm{IPE_B}$ is 0.40 and 0.40. This emphasises the divergence of the standardised correlation coefficients and highlights the sensitivity of IPE to the way in which components are integrated.

Table 4. Integrated assessment of experimental data series.

	IPE values			Rank		
Data series	$\mathrm{IPE_A}$	$\mathrm{IPE_B}$	$\mathrm{IPE_C}$	$\mathrm{IPE_A}$	$\mathrm{IPE_B}$	$\mathrm{IPE_C}$
Naive $(t+1)$	0.04	0.04	0.04	2	2	2
Naive $(t+4)$	0.15	0.40	0.36	6	6	5
Regression $(t+1)$	0.04	0.04	0.04	1	1	1
Regression $(t+4)$	0.14	0.40	0.36	5	5	6
Scaled (low)	0.41	0.41	0.40	9	7	8
Scaled (high)	0.83	0.83	0.89	11	12	12
Bias (low)	0.43	0.43	0.36	10	8	7
Bias (high)	0.87	0.87	0.82	12	12	11
Noise (low)	0.09	0.14	0.14	3	3	3
Noise (high)	0.18	0.46	0.41	7	9	9
Scaled Noise (low)	0.10	0.21	0.25	4	4	4
Scaled Noise (high)	0.20	0.54	0.54	8	10	10

Table 4 also presents some interesting differences when moving from $\mathrm{IPE_B}$ to $\mathrm{IPE_C}$. The latter contains a comprehensive set of orthogonal error measures, and, although there appear to be only minor changes in IPE scores, two things should be noted. First, IPE values range from 0 (for a perfect model) to 1 (for the worst model), so small absolute changes in IPE score (such as 0.43 to 0.36 for Bias (low)) can represent a significant shift in individual overall scoring. Second, the associated rankings of the data series relative to one another can also change when switching from $\mathrm{IPE_B}$ to $\mathrm{IPE_C}$ – notably in the lower half of the scorings. The four top rankings in contrast remained unchanged. This means that an IPE-integrated assessment is both metric and model dependent. Selection of either will control the final tally, and, if it is to be of greater applicability (for example to support cross-study analysis), meaningful, benchmarking operations are required.

The final point to note with this set of results is that, despite having the same (identical by design, as described previously) RMSE, IPE scores for scaled error and bias error are different. This is primarily due to differences in PEP, such that a single local assessment statistic is the controlling item. In so doing, it provides a cautionary justification for a combined error measure such as IPE that can be used to tease out the differences among apparently equivalent models in a process that could easily be perverted.

6 Standardising IPE using naive model benchmarks

Far greater value can be gained from error metrics if they are presented in such a way that they offer a degree of transferability across different catchments or case studies. This is a particularly challenging problem due to the heterogeneity experienced in the hydrological responses of different catchments, located in different physiographic and climatological settings, across different periods. However, one can move towards this goal by benchmarking metrics to a common baseline model, which is applied irrespective of the case study that is of interest. This ensures that the metric is derived by assessing each model's performance relative to a common, simplistic model type that is well understood. It results in additional transferability across different catchments that have a high degree of commonality in their characteristic hydrological responses, because metric values for each catchment indicate the relative, additional performance that is gained over a common baseline modelling approach. It also has the added benefit of demonstrating the complexity of the modelling problem presented by each catchment, as only marginal increases in model performance over the benchmark would indicate a simple problem that warranted a simple modelling solution.

So far, IPE has used the worst performing statistic from the suite of error models under evaluation as the basis for standardising its individual metrics (scaling to one for the worst model, and to zero for a perfect model). Thus, model performance rankings may differ depending on each particular combination of selected metrics adopted and the suite of models included. This arbitrariness is not common hydrological practice, whereby a benchmark model is usually defined a priori, and is independent of comparator models. NSE, for example, compares model performance against a primitive model, comprising the mean of the observed discharge time series as output at all points. IPE model skill is, in contrast, evaluated against a moving target – something that changes according to the mix of models involved, such that reported numerical findings cannot be transferred to other studies.

Standardisation of metrics to a baseline model raises the question of which model to use as the baseline. Clearly, this will vary according to the context of the modelling problem of interest. One possibility is to use a simple linear model benchmark, obtained from least squares linear regression, for the purposes of assessing the extent to which a particular problem is linear or near-linear and so does not require a complex non-linear modelling solution (Abrahart and See, 2007; Mount and Abrahart, 2011). However, in the context of river forecasting models, Seibert (2001) highlights the potential of a simple naive model: "*Obviously, there are more rigorous benchmarks that can be used ... We can also use the observed runoff, shifted backwards by one or more time steps. In this case, we use the observed runoff at time step t as a prediction of the runoff at time step $t + n$. This type of benchmark is especially suitable for forecast models.*"

In this context, the naive model can be thought of as the basic benchmark model "type". However, within this "type", different instantiations of the lag are possible, and the selection will be governed by the nature of the specific modelling problem and data availability.

The adoption of a naive model comparator in hydrological modelling evaluation metrics has a strong linage; it forms a fundamental part of PI (Kitanidis and Bras, 1980). It equates to a one parameter "no knowledge" or "no change situation" model in which the underlying process that is being modelled is assumed to be a Wiener process (i.e. variance increases linearly with time, such that the best estimate for the future is given by the latest measurement). The adoption of a naive $t+n$ model has two key benefits for IPE. First, the comparator model can be easily developed for any river forecasting application. Second, the underlying model will be consistent from catchment to catchment in the respect that it is not controlled by one or more fitted coefficients. The need for n to be consistent and determined by each case study in question is axiomatic.

In the standard application of IPE, the denominator of each component in the IPE equation is the maximum or minimum metric value achieved across a set of models, and is unlikely to ever be zero across all models in the set. However, if a single, standard benchmark model is used, the characteristics of that model fit may result in a zero value for certain metrics. For example, an unbiased benchmark will always have a ME score of zero. Similarly, a naive model will always result in a PEP score of zero. In adapting IPE to a standardised benchmark, this potential issue must be considered. The resultant action should be to either select a benchmark model that will not result in zero values for any of the metrics in the IPE equation, or to adjust the metrics included in the equation so that those evaluating to zero against the chosen benchmark are omitted.

IPE_A and IPE_B cannot be recommended for naive model standardisation since they contain arbitrary and possibly redundant component metrics. IPE_A could also generate scores that exceed unity, whilst IPE_C would encounter a division by zero error in the case of PEP, which will always produce a zero if a naive $t+n$ benchmark model is included. However, given that IPE_C was constructed by means of analytical methods and represents an algorithm structured according to explanatory power, and PEP was the least influential input in IPE_C, PEP could simply be dropped from the equation to produce another variant, IPE_D, thereafter calculated using the four remaining measures (and consequently the overall weighting factor is 0.25, not 0.2):

$$\text{IPE}_\text{D} = \left[0.25\left(\left(\frac{\text{RMSE}_i}{\max(\text{RMSE})}\right)^2 + \left(\frac{\text{RSqr}_i - 1}{\min(\text{RSqr}) - 1}\right)^2 + \left(\frac{\text{ME}_i}{\max|\text{ME}|}\right)^2 + \left(\frac{\text{PI}_i - 1}{\min(\text{PI}) - 1}\right)^2\right)\right]^{1/2} \tag{10}$$

IPE_D will therefore be studied in which:

1. IPE_DW uses each "worst case" individual statistic as a benchmark (as before).

Table 5. $\mathrm{IPE_D}$ analysis of experimental data series.

	IPE values			Rank		
Data series	$\mathrm{IPE_{DW}}$	$\mathrm{IPE_{D1}}$	$\mathrm{IPE_{D4}}$	$\mathrm{IPE_{DW}}$	$\mathrm{IPE_{D1}}$	$\mathrm{IPE_{D4}}$
Naive ($t+1$)	0.04	1.00	0.19	2	2	2
Naive ($t+4$)	0.40	10.37	1.00	8	6	6
Regression ($t+1$)	0.04	0.87	0.14	1	1	1
Regression ($t+4$)	0.40	9.98	0.85	7	5	5
Scaled (low)	0.37	60.59	12.76	5	9	9
Scaled (high)	0.85	165.01	26.68	11	11	11
Bias (low)	0.37	62.36	13.26	6	10	10
Bias (high)	0.87	167.63	27.64	12	12	12
Noise (low)	0.14	3.46	0.38	3	3	3
Noise (high)	0.45	12.48	1.10	9	7	7
Scaled Noise (low)	0.21	5.86	0.83	4	4	4
Scaled Noise (high)	0.53	16.53	1.34	10	8	8

2. $\mathrm{IPE_{D1}}$ uses the naive one-step-ahead prediction as the basis for standardisation (Naive ($t+1$)).

3. $\mathrm{IPE_{D4}}$ uses the naive four-step-ahead prediction as the basis for standardisation (Naive ($t+4$)).

The results of this analysis are provided in Table 5. In this table, the benchmark statistics are used to define the worst case scenario against which everything is measured and standardised. For $\mathrm{IPE_{D1}}$ and $\mathrm{IPE_{D4}}$, we are measuring performance against a naive baseline – any data series that performs worse than these benchmark solutions can be considered particularly poor.

Table 5 presents some interesting results using each of the three benchmarked measures of $\mathrm{IPE_D}$. It depicts similar rankings to those presented earlier for $\mathrm{IPE_A}$, $\mathrm{IPE_B}$ and $\mathrm{IPE_C}$, with the best four and worst two data series being ranked in the same position. In this case, the Regression ($t+1$) and Naive ($t+1$) models are consistently the strongest performing data series assessed by $\mathrm{IPE_{DW}}$, $\mathrm{IPE_{D1}}$ and $\mathrm{IPE_{D4}}$; Scaled (high) and Bias (high) are consistently the worst.

In each scenario, IPE scores for our scaled and bias series are quite different, but IPE scores for Scaled (low) and Bias (low), and for Scaled (high) and Bias (high), are nevertheless similar. This, doubtless, is a reflection of dropping PEP. While there is some difference between these scores, and some of the rankings change as a consequence, there is an argument for modifying IPE to better differentiate between such errors when evaluating models.

Using the naive one-step-ahead model (Naive ($t+1$)) as the baseline ($\mathrm{IPE_{D1}}$) identifies some problems with this particular choice. In this case, only the Regression ($t+1$) has an IPE score less than unity. Having scores that are no longer confined to a common upper range potentially loses something useful. This analysis also highlights the significance of selecting an appropriate benchmark with which to evaluate all other models. In this case, the naive one-step-ahead model would be an inappropriate option as a benchmarking threshold for rejecting models that predict with a longer lead time (such as Regression ($t+4$)).

The benchmark, against which models are evaluated, should be chosen with the same lead time; otherwise, the test is "unfair" and not a true reflection of the accuracy of the models under scrutiny. With this point in mind, a more appropriate baseline might be to use the naive four-step-ahead model (Naive ($t+4$)) – represented as $\mathrm{IPE_{D4}}$. In this case, the simple regression models (Regression ($t+1$) and Regression ($t+4$)), the naive one-step-ahead model (Naive ($t+1$)), and the Noise (low) and Scaled Noise (low) series all perform better than the baseline. However, in this case, it would be wrong to assess the performance of the Regression ($t+1$) and Naive ($t+1$) models against this benchmark as they have a shorter lead time and are thus not facing a "fair" test. The other data series presented all have IPE scores greater than unity so all perform worse than our simple four-step-ahead naive model.

It is also possible to turn this argument on its head; if $t+n$ is seen as a sliding scale, it is possible to offer a series of degraded benchmarks that can be used to quantify the moment at which a particular series crosses a particular threshold (i.e. to establish that the model under test is no better than a $t+n$ naive prediction). This form of assessment may offer rewards in model development operations since the "no change scenario" offers a severe challenge for non-empirical modelling solutions in which the major outcome is greater scientific understanding and not necessarily higher prediction accuracy.

The relative order of the rankings in Table 5 is also worthy of comment. For the best (those ranked in the top four each time) and worst (those ranked 11th and 12th each time) performing data series, there is no change in their relative position from one baseline to the next. However, this is not the case for their absolute IPE scores. For example, the

Regression ($t+1$) model is ranked first for all three baselines, although its IPE scores range from 0.04 (for IPE_{DW}) to 0.87 (for IPE_{D1}). These results emphasise the fact that IPE can provide a useful relative measure of performance within a study, but, to be applicable across studies, a common benchmark must be defined in terms of something meaningful.

A final aspect of IPE that is worthy of further consideration is the role of weights. In the equations presented here, each error measure used in each IPE is equally weighted. This does not have to be the case as more emphasis can be placed on individual components depending on the nature of the modelling requirements. One aspect of IPE that has not been analysed in any study to date is the ability of the modeller to influence the IPE outcome by varying the weighting given to each component metric in the equation. A full analysis of the impact of weight variation is beyond the scope of this study as it, inevitably, will relate to the specific patterns of error in each hydrological model, and this, in turn, will reflect the hydrological modelling challenge of interest. Indeed, we identify this as a potentially worthwhile direction for future work. However, even without a detailed study, it is possible to make some general comments about the impact and potential of using different weights in an IPE. For example, different metrics emphasise different aspects of an error distribution, meaning that specific magnitudinal or timing-related assessment could be of direct operational relevance (e.g. mean absolute error at lead times of 1 to 5 days for river level flood forecasting on the Lower Mekong River; Nguyen and Chua, 2011). Indeed, the squaring of the error value in RMSE will result in a greater emphasis on the model fit at peak flows and in predicting large flood events. For many river flow forecasting problems, this is of primary concern, and it may, therefore, be appropriate to increase the weighting of this metric in the IPE equation. By direct contrast, the use of a relative metric such as MARE will place greater emphasis on model fit at low flows. For drought and low-flow modelling applications, where water shortages and increased pollutant concentration are of interest, it may be appropriate to increase the weight of MARE in the IPE equation. However, not all modelling development activities have an operational focus, and other metrics may be of greater importance to identify systematic problems in model function. To this end, ME (bias) and PI (timing) may be individually weighted as a means of elucidating the sensitivity of a model to these systematic error types.

7 Conclusions

This paper has presented an evaluation of the newly introduced composite index for assessment of model performance known as ideal point error. IPE provides a single point alternative to multiple, possibly contradictory, error measures. The discussion has addressed key issues associated with the use of IPE in the context of river forecasting. The essence of IPE is standardisation of measured error statistics relative to some agreed set of end markers: the selection of a suitable point of reference is a key factor as well as the constituent error metrics. Originally, this was established as the worst performing model in the suite of models under scrutiny. However, in such cases, IPE equates to a moving target that is dependent on the model combination used. Hence, results and conclusions drawn from the analysis are unique to each set of models used in calculating IPE. A more generic use of IPE has been discussed in which a naive $t+n$ step-ahead model is employed for benchmarking purposes. A simple linear model, such as the regression model adopted in this study, could also be used as a more sophisticated benchmark. However, extending the benchmark to ever more sophisticated levels would make cross-comparisons between studies difficult as there is no guarantee the benchmark was being equally derived or applied in each case. Basing the benchmark on one or more naive $t+n$ step-ahead predictions provides a recognised standard that can be consistently applied across different studies, for broader model evaluation purposes.

An area of further work is to examine the interplay between the different errors introduced in this paper and their performance as measured by different error statistics (examining further the themes discussed by Hall, 2001). For example, scaled and bias errors were introduced to the observed record in this study with equal RMSE. In some cases, an integrated IPE provided reasonable differentiation between these errors, in other cases less so. The real-world hydrological relationship between errors and residuals, the latter expressed in terms of theoretical structures and distributions, when applied to data sets with different characteristics, could also be explored.

Edited by: F. Laio

References

Abrahart, R. J. and See, L. M.: Neural network modelling of non-linear hydrological relationships, Hydrol. Earth Syst. Sci., 11, 1563–1579, doi:10.5194/hess-11-1563-2007, 2007.

Abrahart, R. J., Mount, N. J., Ab Ghani, N., Clifford, N. J., and Dawson, C. W.: DAMP: A protocol for contextualising goodness-of-fit statistics in sediment-discharge data-driven modelling, J. Hydrol., 409, 596–611, 2011.

Abrahart, R. J., Antcil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., Shamseldin, A. Y., Solomatine, D. P., Toth, E., and Wilby, R. L.: Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting, Progr. Phys. Geogr., 36, 480–513, 2012a.

Abrahart, R. J., Dawson, C. W., and Mount, N. J.: Partial derivative sensitivity analysis applied to neural network forecasting, Proceedings 10th International Conference on Hydroinformatics, 14–18 July 2012, Hamburg, Germany, 2012b.

American Society of Civil Engineers: Criteria for evaluation of watershed models, J. Irrig. Drain. Eng.-ASCE, 119, 429–442, 1993.

Babovic, V.: Data mining in hydrology, Hydrol. Process., 19, 1511–1515, 2005.

Beran, M.: Hydrograph Prediction – How much skill?, Hydrol. Earth Syst. Sci., 3, 305–307, doi:10.5194/hess-3-305-1999, 1999.

Beven, K.: Prophesy, reality and uncertainty in distributed hydrological modelling, Adv. Water Resour., 16, 41–51, 1993.

Beven, K.: Equifinality and uncertainty in geomorphological modelling, in: The Scientific Nature of Geomorphology, edited by: Rhoads, B. L. and Thorn, C. E., Wiley, Chichester, 289–313, 1996.

Beven, K.: How far can we go in distributed hydrological modelling?, Hydrol. Earth Syst. Sci., 5, 1–12, doi:10.5194/hess-5-1-2001, 2001.

Chiew, F. H. S. and McMahon, T. A.: Assessing the adequacy of catchment streamflow yield estimates, Aust. J. Soil Res., 31, 665–680, 1993.

Cloke, H. L. and Pappenberger, F.: Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures, Meteorol. Appl., 15, 181–197, 2008.

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

Dawson, C. W. and Wilby, R. L.: Hydrological modelling using artificial neural networks, Prog. Phys. Geogr., 25, 80–108, 2001.

Dawson, C. W., Harpham, C., Wilby, R. L., and Chen, Y.: Evaluation of artificial neural network techniques for flow forecasting in the River Yangtze, China, Hydrol. Earth Syst. Sci., 6, 619–626, doi:10.5194/hess-6-619-2002, 2002.

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

Dawson, C. W., Abrahart, R. J., and See, L. M.: HydroTest: further development of a web resource for the standardised assessment of hydrological models, Environ. Model. Softw., 25, 1481–1482, 2010.

Domínguez, E., Dawson, C. W., Ramírez, A., and Abrahart, R. J.: The search for orthogonal hydrological modelling metrics: a case study of 20 monitoring stations in Colombia, J. Hydroinform., 13, 429–442, 2011.

Elshorbagy, A., Panu, U. S., and Simonovic, S. P.: Performance evaluation of artificial neural networks for runoff prediction, J. Hydrol. Eng.-ASCE, 5, 243–261, 2000.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 1: Concepts and methodology, Hydrol. Earth Syst. Sci., 14, 1931–1941, doi:10.5194/hess-14-1931-2010, 2010a.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 2: Application, Hydrol. Earth Syst. Sci., 14, 1943–1961, doi:10.5194/hess-14-1943-2010, 2010b.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, 2009.

Hall, M. J.: How well does your model fit the data?, J. Hydroinform., 3, 49–55, 2001.

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

Kitanidis, P. K. and Bras, R. L.: Real-Time Forecasting With a Conceptual Hydrologic Model: 2. Application and Results, Water Resour. Res., 16, 1034–1044, 1980

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, doi:10.5194/adgeo-5-89-2005, 2005.

Legates, D. R. and McCabe, G. J.: Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 223–241, 1999.

Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. P.: Methods used for the development of neural networks for the prediction of water resources variables: Current status and future directions, Environ. Model. Softw., 25, 891–909, 2010.

Masmoudi, M. and Habaieb, H.: The performance of some real-time statistical flood forecasting models seen through multicriteria analysis, Water Resour. Manage., 7, 57–67, 1993.

Minns, A. W. and Hall, M. J.: Artificial neural networks as rainfall-runoff models, Hydrolog. Sci. J., 41, 388–417, 1996.

Mount, N. J. and Abrahart, R. J.: Discussion of "River flow estimation from upstream flow records by artificial intelligence methods" by M. E. Turan, M. A. Yurdusev [J. Hydrol., 369, 71–77, 2009], J. Hydrol., 396, 193–196, 2011.

Moussa, R.: When monstrosity can be beautiful while normality can be ugly: assessing the performance of event-based flood models, Hydrolog. Sci. J., 55, 1074-1084, 2010.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models 1: A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Nguyen, P. K.-T. and Chua, L. H.-C.: The data-driven approach as an operational real-time flood forecasting model, Hydrol. Process., doi:10.1002/hyp.8347, in press, 2011.

Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrol. Earth Syst. Sci., 13, 999–1018, doi:10.5194/hess-13-999-2009, 2009.

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

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

Weglarczyk, S.: The interdependence and applicability of some statistical quality measures for hydrological models, J. Hydrol., 206, 98–103, 1998.

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

Willems, P.: Model uncertainty analysis by variance decomposition, Phys. Chem. Earth, 42–44, 21–30, 2012.