Hydrol. Earth Syst. Sci. Discuss., 9, 1671–1698, 2012
www.hydrol-earth-syst-sci-discuss.net/9/1671/2012/
doi:10.5194/hessd-9-1671-2012

Hydrology and Earth System Sciences Discussions

Ideal point error for model assessment

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

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

Received: 13 November 2011 – Accepted: 23 December 2011 – Published: 6 February 2012

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

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

When analysing the performance of hydrological models, 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, IPE relies on the adoption of a consistent and recognised benchmarking system. This paper examines one potential option for benchmarking IPE: the use of "persistence scenarios".

1 Introduction

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. It is, therefore, vital that researchers provide adequate clarification of what the specific values of different performance measures really mean in the context of their models. This task is made particularly complex by 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), and that a host of different model evaluation metrics could be applied to a particular solution (Elshorbagy, et al., 2000). Indeed, there is often little consistency in how they are adopted from one study to another (Legates and McCabe, 1999), and some have argued that the choice

HESSD
9, 1671–1698, 2012

of evaluation metrics used is often simply a result of the provision of such 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.

Gupta et al. (2009) proposed a three dimensional combinatorial metric that delivered a dimensionless coefficient: the "Kling–Gupta Efficiency Index" (KGE). This metric represents an evolution of the long-established Nash-Sutcliffe Index (Nash and Sutcliffe, 1970), and delivers a measure of Euclidean distance from a point of ideal error, based upon the model's deviation from the mean and standard deviation of the observed data series. Their metric can be calculated using either un-weighted or re-scalable equations; offering the potential to fine-tune the metric so that it responds more or less strongly to different error types. More recently, Elshorbagy et al. (2010a, b) proposed the Ideal Point Error (IPE). This metric builds upon Gupta et al.'s idea of quantifying the distance from an ideal point of error. However, it is based upon the deviation of a model's multiple goodness-of-fit metrics from their "perfect" scores, rather than statistical measures of deviation. This arguably results in a more flexible evaluation tool that can integrate a wider range of metrics and that makes no assumptions about the statistical distributions of error in a given model. IPE delivers one composite index that can be used as a standalone assessment of model performance, or as a supplemental measure which could support the interpretation of other modelling statistics and/or be of help during a visual inspection of hydrograph error plots. However, limited discussion and evaluation of selection and integration procedures were provided in the original article resulting in numerous unanswered questions about which metrics to choose and how the IPE output is impacted by the particular composition and distribution of the errors in the suite of models under test.

HESSD
9, 1671–1698, 2012

In response, this paper undertakes a more detailed assessment of the limitations and opportunities associated with IPE-based model evaluation 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);

3. suggesting an additional variant of IPE that uses a standardisation approach in which the metrics for each model are evaluated relative to a consistent, naive model benchmark.

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 are employed in this paper: Root Mean Squared Error (RMSE), Mean Average Relative Error (MARE), Mean Error (ME), correlation coefficient (R), R-squared (RSqr), Persistence Index (PI), Percentage Error in Peak (PEP) and Coefficient of Efficiency (CE).

The remainder of this paper is structured as follows: Sect. 2 provides an introduction to IPE and how its different variants are calculated. It also identifies some potential pitfalls in constructing an IPE. Section 3 introduces a set of numerical experiments that were used to evaluate the output consistency of the variants of IPE. Section 4 discusses the error statistics of the numerical experiments and goes on to discuss the identification of, and removal of, strongly correlated metrics within an IPE. Section 5 presents the results of the numerical experiments and provides a discussion and interpretation of their meaning. Section 6 presents and evaluates an IPE standardisation approach based on naive model benchmarking. Section 7 summarises the findings and maps out two potentially rewarding research directions: developing improved theoretical underpinnings and applying preferentially weighted combinations.

HESSD
9, 1671–1698, 2012

2 Ideal point error

2.1 Metric standardisation

IPE is a dimensionless composite index which 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 dataset, 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) published an IPE index that integrated four popular metrics (in which they referred to ME as Mean Bias, MB):

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

for model i, where 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 among the various modelling approaches being compared with regard to their prediction accuracy.

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).

HESSD
9, 1671–1698, 2012

It should be noted that the reported standardisation of the correlation coefficient (R) presented in the original IPE equation (Eq. 1) was not designed to deal with negative scores which could deliver integrated outputs that exceed the maximum upper limit for a perfect score i.e. >1. Equation (2) thus represents an improved variant of the original equation (here termed IPE_B), which includes a more generalised and robust procedure for standardising R that can accommodate its full range [−1,+1]. IPE_B includes a standardised correlation coefficient that ranges from 1 (worst case) to 0 (perfect case) rather than −2 (worst case) to 0 (perfect case) in the original equation. This results in a significant difference in the output of IPE_A and IPE_B, particularly for 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 IPE_A and IPE_B.

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

2.2 The divide by zero problem

The "divide by zero problem" is a computational difficulty for IPE, particularly when the IPE components are replaced by benchmarks (see later discussion). For example, using a naive forecast as a benchmark (i.e. using only antecedent values as the prediction) could well deliver a PEP score of zero. Consequently, the denominator of component S4 will also equate to zero. The same problem could equally apply to other measures for which an optimum value of zero is possible e.g. ME or RMSE. Although such scores are unlikely to occur under standard modelling situations, the issue highlights potential difficulties at the extremities of some metric ranges that the modeller should be aware of.

HESSD
9, 1671–1698, 2012

2.3 Equifinal models

A further potential problem 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 the models are poor because of the manner in which IPE is derived relative to the worst performing model in the suite 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 with this rather than one another and the problem is alleviated. In addition, if IPE produces similar values for different models this would simply highlight the equifinal nature of the models in the suite. In this situation detailed inspection of the corresponding hydrograph might possibly tease out subtle differences between models.

2.4 Metric orthogonality

Dominguez et al. (2011) published a modified IPE index which integrated five popular metrics ordered according to their power of appraisal. It was strongly argued in their paper that the individual statistics that 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 IPE_A metrics were retained in their modified equation (RMSE and ME) and R was replaced by RSqr:

$$IPE_C = \left[0.2\left(\left(\frac{RMSE_i}{\max(RMSE)}\right)^2 + \left(\frac{Rsqr_i - 1}{\min(Rsqr) - 1}\right)^2 + \left(\frac{ME_i}{\max|ME|}\right)^2 + \left(\frac{PI_i - 1}{\min(PI) - 1}\right)^2\left(\frac{PEP_i}{\max|PEP|}\right)^2\right)\right]^{1/2} \quad (3)$$

This IPE was derived from an examination of 22 different statistical metrics for each of 60 models. Principal component analysis (PCA) was used to derive surrogate

HESSD
9, 1671–1698, 2012

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. The analysis led to the five metrics used in Eq. (3) which, it should be noted, is dependent on the dataset involved.

The use of a comprehensive PCA approach to analysing 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 that will bias the IPE output through the identification of metrics whose correlation coefficients are similarly high. 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 include. We, therefore, perform just such an analysis in our evaluation of IPE later in this paper.

3 Numerical experiments

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 cumecs $\times 10^2$ at the site of the Three Gorges Dam, on the Yangtze River in China. The data covers 4 July 1992 to 13 August 1992 and comprises 160 observed records. The data set can be downloaded from the HydroTest website (Dawson et al., 2007, 2010). Further particulars on the origins of the dataset can be found in Dawson et al. (2002). Figure 1 provides a hydrograph of these data.

The IPE variants specified in Eqs. (1–3) are evaluated and benchmarked in this paper using twelve simple data series which are compared against the observed record. New sentence reads "Two versions of different comparators are included, representing large and small deviations from 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.

HESSD
9, 1671–1698, 2012

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.

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 type of error 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)*).

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

where r_n is the regression coefficient for time lag n ($n = 1$; $n = 4$), and k_n is the constant offset ($n = 1$, $n = 4$). 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. Scaled errors that are proportional to the magnitude of the observed flow. These errors can be expressed as:

$$\hat{Q}_i = cQ_i \tag{6}$$

HESSD
9, 1671–1698, 2012

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 errors are referred to as *Scaled (Low)* for $c = 1.25$ and *Scaled (High)* for $c = 1.5$.

4. Bias errors that increment the observed discharge by a constant amount (b) according to the following equation and as such equate to a vertical displacement of the original record:

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

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 are 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 errors are referred to as Bias (Low) and Bias (High) respectively.

5. Errors in which random noise has been added to the observed record.

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

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. In the second case the standard deviation adopted is half that of the standard deviation of the observed record. These values were chosen as they represent a reasonable distribution of noise without generating negative flow values. These errors are referred to as *Noise (Low)* and *Noise (High)* respectively.

6. Errors in which the random noise added to the observed record in Eq. (8) above has been scaled by the square of the observed record. This leads to

HESSD
9, 1671–1698, 2012

proportionally larger errors at high flows and lower errors at low flows.

$$\hat{Q}_i = Q_i + NQ_i^2/k \quad (9)$$

in which k is a value chosen to ensure scaled errors do no lead to negative flows. In this case, setting k to the square of the mean of the observed record (285.82) leads to acceptable results. The two error models are referred to as *Scaled Noise (Low)* and *Scaled Noise (High)* coinciding with the amount of random noise added from Eq. (8) above.

Figure 2 provides error plots of each of these data series compared with observed flow. The figures show similar performance of the naive and regression models with the two models based on one step-ahead prediction demonstrating low errors across the range of the observed record. The scaled errors show, not surprisingly, a linear increase in error as observed flow increases, while the bias errors show consistent error across the same range. The two noise models (low and high) show a reasonably even spread of error across the range of the observed record, while the scaled noise displays heteroscadastic error.

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 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.

HESSD
9, 1671–1698, 2012

There are a few notable points to be made about these results. First, note that both Naive models return PEP values of zero (the best score). This is because both are generated directly as a time-shift of the observed data, and consequently have the same peak score as the observed record. This brings into question individual error measures such as these 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. Secondly, although most of the error statistics return similar results 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 $k_1 = -0.332$ factor 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 results. Once again, this highlights the dangers of using individual measures that may provide results which are contradictory to what is actually being measured.

Another notable point from these results is that despite returning perfect scores for RSqr and R, the scaled and bias errors 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. All these results emphasise that individual error statistics cannot be relied upon to provide an objective measure of model performance. It is only when error statistics are compared or combined that an overall picture of model performance emerges.

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 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

HESSD
9, 1671–1698, 2012

study the data series have been artificially generated and, as a consequence, a number of the data series present near identical results according to many of the metrics. 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 IPE_C as they are closely related to RMSE which would be retained. Retaining 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 this case, because 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 IPE_C which is based on a sound hydrological analysis.

4.3 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 IPE_A to 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 IPE_A and IPE_B. The IPE_A and 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: IPE_A is 0.15 and 0.14 respectively; 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 also presents some interesting differences when moving from IPE_B to IPE_C which consists of an orthogonal and comprehensive set of error measures. Although

HESSD
9, 1671–1698, 2012

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 the Bias (Low) error) 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 IPE_B to 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 RMSE (as defined in the previous section), 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 a controlling item. In so doing it provides a cautionary justification for a combined error measure such as IPE which can be used to tease out the differences among apparently equivalent models in a process that could easily be perverted.

5 Standardising IPE using naive model benchmarks

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 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. CE (Nash and Sutcliffe, 1970), for example, compares model performance against a primitive model, comprising the mean of the observed discharge time series of the calibration period 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.

HESSD
9, 1671–1698, 2012

In order for IPE to be applicable across real-time, flood forecasting studies, IPE could also be standardised using a simple naive model benchmark. A number of candidate baselines are available from the models introduced in this study. One such benchmark is the naive ($t+n$) model which simply predicts the current value using a value recorded at n previous time steps (a persistence index baseline). The need for n to be consistent and determined by each case study in question is axiomatic. It is also possible to provide 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).

The IPE_A and IPE_B variants cannot be adapted for naive model standardisation since IPE_A generates scores that exceed unity, whilst IPE_B 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. IPE_C also uses PEP and so should be similarly discarded. However, given that it was constructed by means of analytical methods and represents an algorithm structured according to explanatory power, and PEP is the least influential input in IPE_C, PEP could be dropped from the equation to produce the variant IPE_D; thereafter calculated using the four remaining measures (and consequently the overall weighting factor is 0.25, not 0.2):

$$IPE_D = \left[0.25\left(\left(\frac{RMSE_i}{\max(RMSE)}\right)^2 + \left(\frac{Rsqr_i - 1}{\min(Rsqr) - 1}\right)^2 + \left(\frac{ME_i}{\max|ME|}\right)^2 + \left(\frac{PI_i - 1}{\min(PI) - 1}\right)^2\right)\right]^{1/2} \quad (10)$$

IPE_D will therefore be studied in which:

1. IPE_{DW} uses each "worst case" individual statistic as a benchmark (as before).
2. IPE_{D1} uses the naive one-step-ahead prediction as the basis for standardisation (Naive ($t+1$)).
3. IPE_{D4} uses the naive four-step-ahead prediction as the basis for standardisation (Naive ($t+4$)).

HESSD
9, 1671–1698, 2012

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 IPE_{D1} and IPE_{D4} we are measuring performance against a naive baseline – any data series which 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 IPE_D. It depicts similar rankings to those presented earlier for IPE_A, IPE_B and 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 of the data series assessed for IPE_{DW}, IPE_{D1} and IPE_{D4}; and Scaled (High) and Bias (High) errors are consistently the worst.

Although the IPE scores for each of the baselines are quite different for each of the scaled and bias errors, IPE scores of Scaled (Low) and Bias (Low); and Scaled (High) and Bias (High); are similar for each baseline. 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 some 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 (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 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 are predicting 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 IPE_{D4}. In

HESSD
9, 1671–1698, 2012

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) errors are all performing 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 are all worse than this simple case.

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.

6 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

HESSD
9, 1671–1698, 2012

point alternative to multiple, possibly contradictory, error measures. The discussion has addressed some of the issues associated with the use of IPE. The essence of IPE is standardisation of measured error statistics relative to some agreed set of end markers: such that 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 which is dependent on the model combination used. Hence, results and conclusions drawn from the analysis are unique to the 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 adopted for benchmarking purposes. A simple linear model, such as the regression model used 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.

Another area of further work is to explore the relative weightings of individual statistics within an IPE. 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

HESSD
9, 1671–1698, 2012

modelling requirements. For example, when evaluating water resources models, an IPE that places more emphasis on low flow statistics (such as MSRE, or MSLE) may be preferable.

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.

Beven, K.: Prophesy, reality and uncertainty in distributed hydrological modelling, Adva. 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.

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. Geog., 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. Modell. 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. Modell. Softw., 25, 1481–1482, 2010.

Domínguez, E., Dawson, C. W., Ramírez, A., and Abrahart, R. J.: The search for orthogo-

HESSD
9, 1671–1698, 2012

nal 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.
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.
Hall, M. J.: How well does your model fit the data?, J. Hydroinform., 3, 49–55, 2001.
Masmoudi, M. and Habaieb, H.: The performance of some real-time statistical flood forecasting models seen through multicriteria analysis, Water Resour. Manag., 7, 57–67, 1993.
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 (2009) 71–77], J. Hydrol., 396, 193–196, 2011.
Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models 1: A discussion of principles, J. Hydrol., 10, 282–290, 1970.
Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, 2007.
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. Modell. Softw., 24, 311–321, 2009.
Willems, P.: Model uncertainty analysis by variance decomposition, Phys. Chem. Earth, in press, doi:10.1016/j.pce.2011.07.003, 2012.

HESSD
9, 1671–1698, 2012

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

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

HESSD
9, 1671–1698, 2012

Table 2. Error statistics for each data series ("best" result in bold, "worst" result in italic for each statistic).

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*

HESSD
9, 1671–1698, 2012

Table 3. Cross correlation between metrics based on the 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

HESSD
9, 1671–1698, 2012

Table 4. Integrated assessments of error models.

	IPE Values			Rank		
Data series	IPE_A	IPE_B	IPE_C	IPE_A	IPE_B	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

HESSD
9, 1671–1698, 2012

Table 5. Results for IPE_D analysis of data series.

Data Series	IPE Values			Rank		
	IPE_{DW}	IPE_{D1}	IPE_{D4}	IPE_{DW}	IPE_{D1}	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

HESSD
9, 1671–1698, 2012

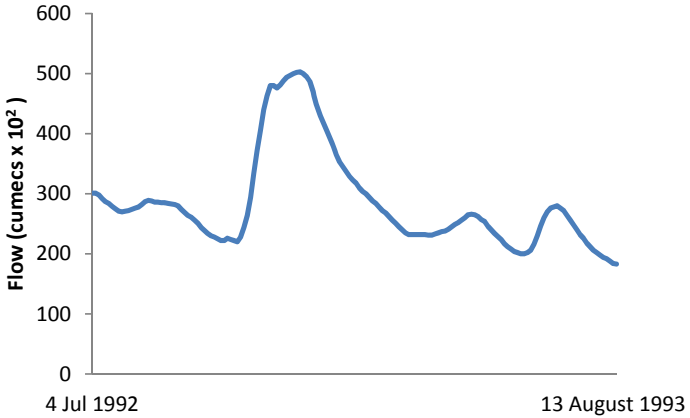

Fig. 1. Hydrograph of observed flow from Three Gorges Dam, Yangtze River, China.

HESSD
9, 1671–1698, 2012

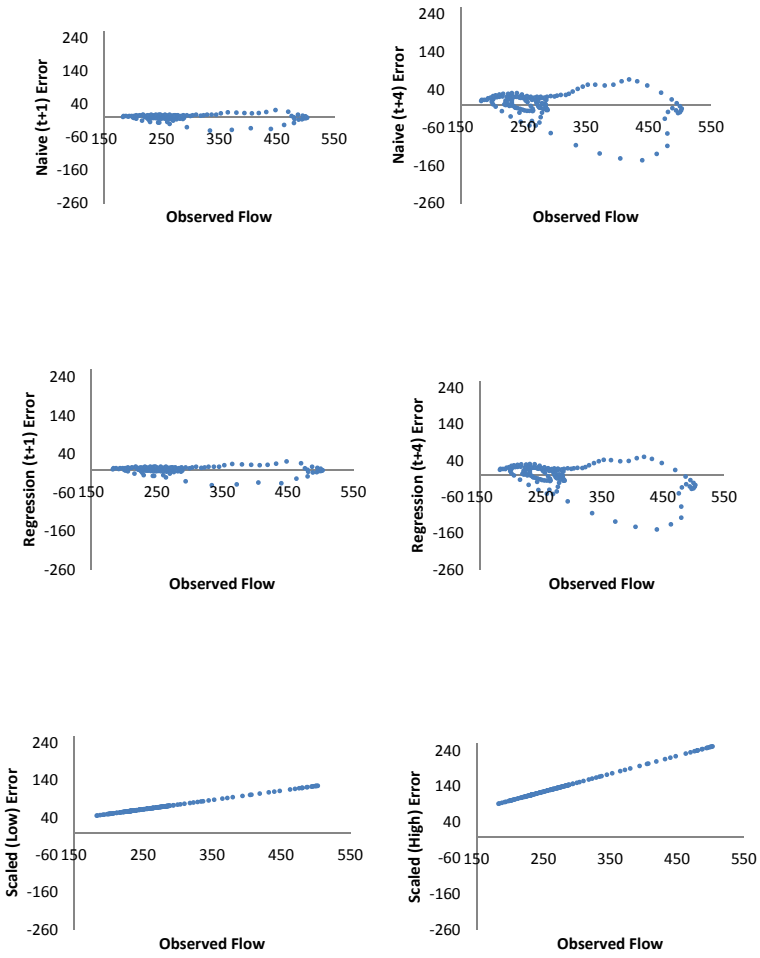

Fig. 2. Caption on next page.

HESSD
9, 1671–1698, 2012

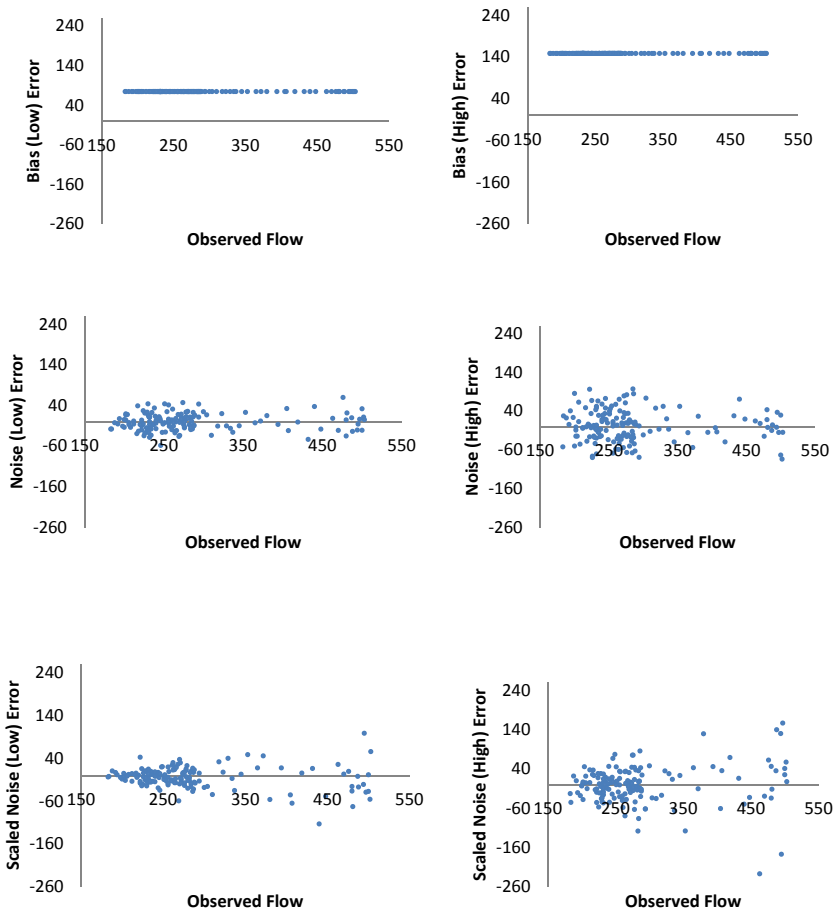

Fig. 2. Error plots of each data series with respect to observed flow (measurements in cumecs $\times 10^2$).

HESSD
9, 1671–1698, 2012