HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts

Dawson, C.W.[1*], Abrahart, R.J.[2] and See, L.M. [3]

[1] Department of Computer Science, Loughborough University, Loughborough, LE11 3TU, UK
[2] School of Geography, University of Nottingham, Nottingham, NG7 2RD, UK
[3] School of Geography, University of Leeds, Leeds, LS2 9JT, UK

Abstract

This paper presents details of an open access web site that can be used by hydrologists and other scientists to evaluate time series models. There is at present a general lack of consistency in the way in which hydrological models are assessed that handicaps the comparison of reported studies and hinders the development of superior models. The HydroTest web site provides a wide range of objective metrics and consistent tests of model performance to assess forecasting skill. This resource is designed to promote future transparency and consistency between reported models and includes an open forum that is intended to encourage further discussion and debate on the topic of hydrological performance evaluation metrics. It is envisaged that the provision of such facilities will lead to the creation of superior forecasting metrics and the development of international benchmark time series datasets.

Key words

Hydrological models, performance testing, evaluation metrics

Software availability
Name of the product: HydroTest website
Developed by: Christian W. Dawson
Contact address: Department of Computer Science, Loughborough University, Loughborough, LE11 3TU, UK.
Tel.: +44-1509-222684;
Fax.:+44-1509-211586;
E-mail: C.W.Dawson1@hydrotest.org.uk
Available since: 2005
Coding language: PHP: Hypertext Pre-processor
Availability: via http://www.hydrotest.org.uk
Cost: Free

* Corresponding author

1. Introduction

Krause *et al*. (2005: p89) list three main reasons for the performance assessment of hydrological models which can be paraphrased as follows: (1) to provide a quantitative indicator of model skill at reproducing catchment behaviour; (2) to provide a means for evaluating improvements to the model or method of modelling; (3) to provide a mechanism for comparing the results obtained in different modelling studies. To select the best model from a set of competing models available for a particular application is also a difficult undertaking in that the purpose of the model must be matched to the measures that are used to assess forecasting skill with regard to operational requirements. The process of assessing the performance of a hydrological model requires both subjective and objective estimates to be made of the relationship that exists between simulated behaviour and observed behaviour. Evaluation is reported in terms of closeness of fit and in most cases with respect to observations recorded at the catchment outlet. The most fundamental method of assessing model performance in terms of functional behaviour is through a visual inspection of the differences between simulated and observed time series plots. In doing so, the hydrologist can formulate subjective assessments of the model behaviour with respect to systematic (e.g., overprediction or underprediction) and dynamic (e.g., timing, rising limb, falling limb, and base flow) behaviour of the model. For objective assessment, however, the hydrologist will require the production of one or more mathematical estimates of the error that occurs between the simulated and observed hydrological time series record.

There is a pressing need to develop superior metrics for the quantitative evaluation of hydrological forecasts. Metrics are defined as a system of parameters, or methods of

quantitative assessment, for something that is to be measured. Metrics define what is to be measured along with the processes that are used to perform such measurement. Teegavarapu and Elshorbagy (2005: p200) in a consideration of traditional error metrics concluded that conventional measures "have limited use and may not always provide a comprehensive assessment of the performance of the model developed for a specific application". Earlier models could be reported in terms of output evaluations that were considered to be "sufficient" or "acceptable" or "adequate" since the inner workings of the model had strong theoretical support. Traditional evaluation metrics were intended to support this level of assessment, but struggled to maintain a realistic distinction between different models, developed on different combinations of structures and parameter sets, and could not be used to validate model structure. Moreover, most calibrated models possess sufficient parameters to produce an "acceptable result" in terms of a "reasonable fit to output discharge", in contrast to the operational requirement which is for solutions that will deliver "best possible accuracies". The last decade has witnessed a virtual explosion in the application of soft computing methodologies throughout the hydrological sciences and much of that effort has been directed towards the production of superior forecasting solutions. Numerous other opportunities are also being explored, such as the potential for significant savings to be made in terms of model development time and effort, or the need to model nonlinear systems where traditional parameter estimation techniques are not convenient (Singh and Woolhiser, 2002). For reviews of neural network applications in the hydrological sciences see: ASCE (2000), Dawson and Wilby (2001), Maier and Dandy (2000). This rapid and widespread uptake of "data-driven technologies" or "smart solutions" is not intended to be a substitute for conceptual watershed modelling. It has nevertheless created a mass of operational solutions that can

produce improved hydrological modelling outputs but which, *ipso facto*, have no theoretical support and where "avoidance of overfitting" is required (Giustolisi and Laucelli, 2005). This field of modelling is based on numerous combinations of computational algorithms and numerical assessment procedures; total reliance is placed upon evaluation metrics throughout both model construction and model verification operations. The modelling power of a "smart solution" has no theoretical constraints and under the right circumstances the final product should be able to surpass the imperfect representational capabilities of traditional evaluation metrics. To develop superior soft computing solutions will in consequence require the development of more efficacious metrics upon which to construct and test the next generation of models that are needed to service the hydrological demands of operational planners and managers. The first step in this process is to bring together the established set of past and present model evaluation procedures in a standardised toolbox. This will promote the detailed exploration and cross-comparison of each metric in terms of (a) potential merits and (b) possible drawbacks. The toolbox will thereafter serve as a springboard for subsequent improvements.

This paper is about the assessment of time series forecasting models. However, commeasure with past (e.g. Clarke, 1973) and present (e.g. Jakeman *et al*., 2006; Refsgaard and Henriksen, 2004; Refsgaard *et al.* 2005) concerns about the need to develop consistent hydrological modelling protocols and terminologies, it is important to be clear about what does or does not constitute a time series forecasting model. A prediction is a statement or a claim that a particular event will occur (the etymology of this word comes from the Latin præ- "before" plus dicere "to say"). It covers something

that is expected to occur provided that a set of preconditions is (or is not) satisfied. It asserts the existence of a temporal relationship between the statement that is being made and the event itself, but the latter does not need to possess a specific temporal component, or the timing of the event can be directional but open ended. Anderson and Burt (1985: p2) adopt a similar stance and state that "the aim of most hydrological modelling is to provide a prognosis of the future performance of a hydrological system; such a judgement may be made with respect to real-time (forecasting) or without specific time reference (prediction)". Klemes (1986) also supports this viewpoint: "Following the general usage in hydrology, by forecasting is meant real-time forecasting while prediction is used for estimation of future states which are not specified in time…". Thus expectations that involve a specified temporal component can be treated as a special case of prediction; often sub-divided into the estimation of future conditions (forecasting), present conditions (nowcasting) or past conditions (hindcasting). The temporal component that is involved in nowcasting operations remains a matter of debate and the term should perhaps be restricted to short term forecasts of up to about two hours ahead. The essential difference between shorter period nowcasts and longer period forecasts is associated with the rapid decline in accuracies that will occur over time due to the complex and ever changing nature of our hydrological and meteorological environments. The upshot to this rationalisation of time series forecasting models is that models which are designed to produce estimates of future conditions or variables in terms of specified temporal units should be called forecasting models irrespective of the processes or datasets that are used to build and test such tools; it is their intended application that counts - not the methods or the mechanics of their development. The temporal component in real time forecasting requires specified temporal input units and output units and for the model inputs to be

processed with sufficient speed such that no backlog occurs. Thus, sampling intervals and updating procedures are important in an operational setting as is the progressive integration of digital networks with telemetric stations.

Scientific and engineering shortcomings in the 1980s and 1990s led to an assessment of published papers which noted that model developers typically do not provide consistent or standard quantitative evaluation criteria. This does not assist either readers or users in determining how well their model reproduces the observed dataset or how well their model compares with other models (ASCE, 1993). There was, and still is, a general lack of transparency, objectivity and consistency with respect to the manner in which hydrological models are assessed, evaluated and/or compared in reported studies (Legates and McCabe, 1999). Not only are a number of different error metrics used from one study to the next, but there is also no guarantee that in cases where the same error metric is used, it has been applied consistently and correctly. For these reasons it is often difficult to compare objectively the results presented in different studies. Moreover, analogous terms and expressions are being used in different papers, to represent either the same metric or different metrics. This matter is a source of potential confusion, that extends from a simple doubling-up based on replacement expressions and formats (e.g. 'mean' is degraded to 'average' or 'ratio' is converted into a 'percentage'), to the more disturbing use of different terms and labels for the description of identical metrics (e.g. Mean Absolute Relative Error is sometimes shortened to Mean Relative Error; Karunanithi *et al*., 1994; Elshorbagy *et al*., 2000; Teegavarapu and Elshorbagy, 2005), or, in reverse, using identical labels to represent different metrics (e.g. Normalised Root Mean Squared Error is being calculated in a number of different manners; Atiya *et al*.,

1999; Jain and Srinivasulu, 2005; Karunasinghe and Liong, 2006). The need for a universal vocabulary, based on consistent equations, is obvious but until such time as a definitive source has been established there remains a requirement to include each individual algorithm in each individual paper, or for each interested reader to refer to previous papers, in order to obtain detailed particulars and a fuller understanding of the metric that is being reported. However, even cross referencing and the use of citations is not a foolproof method, since previous errors and omissions can be propagated throughout the field e.g. incorrect bracket notation in a published equation (MAE; Chang *et al*., 2002). It is better to be consistent and transparent and to promote collective development activities, that are open to inspection, and that would support and encourage maximum potential testing, correctness and peer group endorsement.

Chiew and McMahon (1993) argued that amongst practitioners the use of different evaluation metrics for hydrological modelling purposes was related to whether or not the relevant procedures were provided in common software packages. Identical stumbling blocks occur in the case of multi-purpose statistical software for data analysis operations that are provided on the web: e.g. StatCrunch (http://www.statcrunch.com/) where usage has grown to the point that it is no longer feasible to offer free access and support. The proposed solution to such problems is the construction of a web processing platform that is dedicated to the evaluation of hydrological forecasting models. This solution will also be used to promote a consideration of metrics that might not otherwise be considered, and present the latest innovations in this field, for operational appraisal and testing purposes. There is, moreover, no universal measure of performance and the metric(s) that is (are) chosen should correspond to the particular needs of each individual application. Further,

since single measurements are insufficient, the use multiple measurements is a common occurrence such that hydrologists must be able to select their preferred metrics from a comprehensive range of potential options. Jakeman *et al*. (2006) note that other criteria should be tested, including tests on residuals for cross-correlation with predictors, heteroscedasticity, autocorrelation etc. It is also desirable that modellers compare the relative performance of different evaluation procedures tested under different conditions; for a recent comparison of nine different evaluation metrics tested on three different types of synthetic error that were introduced into a discharge time series see Krause *et al*. (2005).

The purpose of this paper is to present the main details of an open access web site that can be used by hydrologists and other scientists to evaluate their (hydrological) models – providing a broad spectrum of quantitative tests and consistent measures of performance. The site has been developed primarily to assess hydrological model time series forecasts, but it can also be used to evaluate any other models that produce time series outputs. The aim of this site is to provide a uniform set of quantitative tests that will ensure transparency and consistency between reported studies. By establishing a single web site location for model evaluation purposes we are hoping to provide scientists with the following:

1. Research support;
2. Fast and efficient service;
3. Method for the accreditation of computed results;
4. Evaluation tools built on community involvement and testing;

5. Metadata rule base that highlights and standardises implementation issues;
6. International warehouse of popular and seldom used mathematical algorithms;
7. International forum for open discussion and debate that will further this subject.

In return, users of the site are requested to:

1. Register with the site;
2. Cite the web site in related publications;
3. Participate in transparent discussions and debates on testing procedures;
4. Inform the site developers of any improvements, corrections and amendments.

The privacy of users is important and details provided during registration will not be passed on to any other parties. Registration will allow the site developers to monitor uptake and usage of the site and keep users informed of any updates and corrections.

The remainder of this paper is arranged as follows. Section 2 discusses the evaluation metrics that have been incorporated into the web site. Section 3 discusses the different types of errors produced by hydrological models; four example model outputs are provided and the relationships between the evaluation metrics and the model output errors are discussed. Section 4 provides a case study – showing how the site was used to establish the 'best' model from a series of model calibrations. Section 5 introduces the site itself, its interface, and how it should be used. Section 6 presents some conclusions and discusses future developments for the site and this project.

2. Evaluation metrics

This section will describe each evaluation metric that is calculated on the HydroTest web site. ASCE (1993) reviewed a number of evaluation metrics for the assessment of hydrological models which were split into two distinct categories; those for evaluating continuous hydrographs and those that should be applied to single-event models. HydroTest computes both types of evaluation metric and the user should be aware of the limitations and appropriate applications of each of metric that is reported. The different types of metric on the web site can be categorised as follows:

a. Statistical parameters of observed and modelled time series datasets.
b. Statistical parameters of the residual error between observed and modelled time series datasets.
c. Dimensionless coefficients that contrast model performance with accepted norms or recognised standards.

In the following equations, Q_i is the observed (i.e. expected) value, $\hat{Q}_i$ is the modelled (i.e. forecast) value (where i = 1 to n data points), $\overline{Q}$ is the mean of the observed data, and $\widetilde{Q}$ is the mean of the modelled values.

The metrics are aimed primarily at evaluating river discharge forecasting models (for example, models predicting discharge measured in cumecs) although most of the metrics can also be used to evaluate other forecasting systems – for example, in the estimation of

river stage (e.g. level of water measured in metres), water quality parameters (e.g. CO_2 concentration), water temperature, air temperature, etc. Care must be taken, however, when using these metrics to evaluate other forecasts to ensure that the measurements are applicable. For a more detailed discussion of particular evaluation metrics the interested reader is directed to: Armstrong and Collopy (1992); ASCE (1993); Beran (1999); Green and Stephenson (1986); Hall (2001); Krause *et al*. (2005); or Legates and McCabe (1999). Chiew and McMahon (1993) and Houghton-Carr (1999) provide comparisons of quantitative metrics with qualitative assessments based on expert opinions.

Table 1 provides a list of indicative references for the numerous metrics that are computed on the web site and presented in this paper. Hydrologists are invited to recommend other metrics and case studies for incorporation in subsequent processing operations and reporting activities.

2.1 Statistical parameters of observed and modelled time series datasets

The general characteristics of each individual dataset can be described or summarised using standard descriptive statistics. It must be stressed that the calculation of statistical parameters is not in itself a testing operation since it involves the production of individual descriptors for each dataset and no direct comparison is performed. It does nevertheless lead to various possibilities for subsequent quantitative and qualitative comparisons. The overall level of agreement between the two datasets with respect to their computed statistical parameters can be evaluated in direct terms of 'relative magnitude' or through the use of mathematical constructs such as 'ratios'. Eight individual descriptive statistics are computed for the observed and modelled time series datasets comprising: minimum,

maximum, mean, variance, standard deviation, skewness, kurtosis, and lag-one autocorrelation coefficient. Chiew and McMahon (1993) provided equations for what was regarded to be a set of simple parameters that could be used to describe the characteristics of each particular time series: i.e. mean, standard deviation, skewness, lag-one autocorrelation coefficient. The authors accepted that other statistical parameters such as kurtosis, or higher order moments and higher-lag autocorrelation coefficients, could be used but maintained that since most hydrological applications at that point were based on a comparison of observed and modelled means (and sometimes standard deviations) their four chosen parameters were sufficient to describe the similarities that exist between two discharge time series datasets. However, soft computing methodologies are sensitive to other factors, so four other standard descriptive parameters have also been included on the web site: minimum and maximum to represent range; variance to represent statistical dispersion; kurtosis to cover the shape of the distribution in terms of 'peakedness', with higher values meaning that more of the variance is due to infrequent extreme deviations, as opposed to frequent modest-sized deviations.

2.2 Statistical parameters of residual error between observed and modelled time series datasets

It is important to provide a quantitative assessment of model error expressed in terms of the units of the variables of interest and that can thereafter be interpreted in a meaningful manner. It is also desirable to include in such evaluations a consideration of other factors such as the number of parameters and to test for systematic errors. Measurements in this category are termed *absolute errors*. It is also important to provide a quantitative assessment of model error expressed in terms of unbiased unit-free metrics that can

thereafter be used to support interpretation in a purposeful context. Measurements in this category are termed *relative errors* and record the mismatch that occurs between the observed and modelled values, expressed in terms of ratios and percentages, based on the relative relationship that exists between observed records and model error values. The use of relative error measurements is intended to redress the limitations of absolute error measurements which, although useful, do not necessarily give an indication of the importance of an error. For example, an error of 1cm is very significant over a measurement of 2cm, but virtually irrelevant over a measurement of 10m. The two different types of assessment metric in this section: (a) can be used to provide a comparison between equivalent models, produced on the same catchment, but cross catchment comparisons are invalid since no common standard exists; and (b) possess no fixed criterion in terms of what does or does not constitute a "good" value; it is nonsense to state that "the model is good (bad) because a particular evaluation measure is less (greater) than x", unless one is referring to a specific degree of accuracy that is relevant to a particular forecasting application. For absolute parameters such accuracies must be expressed in identified units of measurement; for relative parameters such accuracies must be expressed in terms of relative proportions.

2.2.1 Estimation of absolute parameters

Equation 1 is used to calculate the Absolute Maximum Error (AME). This metric records in real units the magnitude of the worst possible positive or negative error that the model has produced. It is a non-negative metric that has no upper bound and for a perfect model the result would be zero. It does not attempt to represent in a direct manner the level of overall agreement between the two datasets and individual outliers can have a marked

influence or produce a misleading effect. The measure is nevertheless useful in situations where it is important to establish whether or not a particular environmental threshold has been exceeded i.e. maximum permitted error.

$$\mathrm{AME} = \max(\left|Q_i - \hat{Q}_i\right|) \quad (1)$$

Equation 2 is used to calculate the Peak Difference (PDIFF). This metric records in real units how well the highest output value in the modelled dataset matches the highest recorded value in the observed dataset. It is a *signed* metric that has no upper bound and for a perfect model the result would be zero. As a signed metric this measure indicates whether or not the forecasts are *biased* i.e., does a systematic error exist such that the forecasts tend to be either disproportionately positive or negative. The metric is positive if a model under-estimates the overall actual values, or negative if a model over-estimates the overall actual values. It is worth noting that some authors (for example, Chang et al. 2001) present these kinds of statistics the opposite way round – i.e. negative values for an under-estimate and vice versa. The signed metrics produced by the HydroTest site consistently represent under-estimates as positive values and over-estimates as negative values.

PDIFF does not attempt to represent in a direct manner the level of overall agreement between the two datasets and the temporal relationship that exists between the highest magnitude in each dataset is not considered. There is no requirement for the two peak values to coincide; indeed, the highest peak in the observed dataset might occur near the beginning of the series, whereas the highest peak in the modelled dataset might appear towards the end of the period that is being modelled. PDIFF is nevertheless useful in terms of providing an indication as to whether or not the model is able to produce a

similar range of forecast values to that which occurs in the observed dataset. This being the case it would appear to be more appropriate for single-event modelling as opposed to continuous modelling and care must be taken when applying this metric to continuous hydrographs since the two maximums that are being compared might derive from different storm events in the observed and modelled datasets.

$$\text{PDIFF} = \max(Q_i) - \max(\hat{Q}_i) \quad [\text{for } i = 1 \text{ to } n] \tag{2}$$

Equation 3 is used to calculate the Mean Absolute Error (MAE). This metric records in real units the level of overall agreement between the observed and modelled datasets. It is a non-negative metric that has no upper bound and for a perfect model the result would be zero. It provides no information about under-estimation or over-estimation. It is not weighted towards high(er) magnitude or low(er) magnitude events, but instead evaluates all deviations from the observed values, in both an equal manner and regardless of sign. MAE is comparable to the Total Sum of Absolute Residuals (TSAR; Stephenson, 1979) that was recommended for the comparison of single-event models in a major review of evaluation criteria that was conducted by Green and Stephenson (1986) and thereafter endorsed by ASCE (1993).

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left| Q_i - \hat{Q}_i \right|}{n} \tag{3}$$

Equation 4 is used to calculate the Mean Error (ME). This signed metric records in real units the level of overall agreement between the observed and modelled datasets. It is unbounded and for a perfect model the result would be zero. However, a low score does

not necessarily indicate a good model in terms of accurate forecasts, since positive and negative errors will tend to cancel each other out and, for this reason, MAE (Equation 3) is often preferred to ME.

ME is not weighted towards high(er) magnitude or low(er) magnitude events, but instead evaluates all modelling deviations from the observed values in an equal manner including the sign. To compare ME values across different variables or across events with different magnitudes (for non-negative variables) it can be normalised; ME is divided by the mean of the observed values over the period or event to produce the Normalised Mean Bias Error (NMBE; Jain and Srinivasulu, 2004; 2005).

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(Q_i - \hat{Q}_i) \tag{4}$$

Equation 5 is used to calculate the Root Mean Squared Error (RMSE). This metric records in real units the level of overall agreement between the observed and modelled datasets. It is a non-negative metric that has no upper bound and for a perfect model the result would be zero. It comprises a weighted measure of the error in which the largest deviations between the observed and modelled values contribute the most. It is computed on squared differences and assessment is thus biased in favour of peaks and high(er) magnitude events, that will in most cases exhibit the greatest errors, and be insensitive to low(er) magnitude sequences. It is in consequence more sensitive than other metrics to the occasional large error: the squaring process gives disproportionate weight to very large errors. RMSE can provide a good measure of model performance for high flows (Karunanithi *et al*., 1994), but significant variations in the assessment of different

catchments will occur, since the evaluation metric is dependent on the scale of the dataset that is being analysed. It is perhaps better to report RMSE, rather than Mean Squared Error (MSE; Chang *et al*., 2004; Chen *et al*., 2006; Furundzic, 1998; Riad *et al*., 2004;), because RMSE is measured in the same units as the original data, rather than in squared units, and is thus more representative of the size of a "typical" error. MSE was at one point the most widely used measure of overall accuracy for a forecasting method but it is also the method that has incurred the most criticism (e.g. Clements and Hendry, 1993). RMSE is usually similar in magnitude to, but slightly larger than, MAE (Equation 3) and the extent to which RMSE exceeds MAE is an indicator of the extent to which outliers (or variances in the differences between the modelled and observed values) exist in the datasets (Legates and McCabe, 1999). To compare RMSE values across different variables or across events with different magnitudes (for non-negative variables) it can be normalised; RMSE is divided by the mean of the observed values over the period or event to produce the Relative Root Mean Squared Error ($RMSE_r$; Fernando *et al*., 1998; Pebesma *et al.*, 2005). Jain and Srinivasulu (2004; 2005) refer to this metric as Normalised Root Mean Squared Error (NRMSE). It should also be noted that a number of other formulations exist for calculating Normalised Root Mean Squared Error e.g. Atiya *et al*. (1999) or Karunasinghe and Liong (2006). RMSE is comparable to Sum Squared Error (SSE; Giustolisi and Laucelli, 2005; Lauzon *et al*., 2006) and the latter was once one of the most popular metrics for evaluating hydrological simulation models (Diskin and Simon, 1977). SSE, used for the evaluation of single event models, is termed 'Simple Sum of Squared Residuals' (G); for the evaluation of continuous modelling over a number of events it is termed Total Sum of Squared Residuals (TSSR). SSE measures were also recommended in the review of Green and Stephenson (1986) and thereafter

endorsed by ASCE (1993). It should be noted that G and TSSR, like MSE, are expressed in squared units whereas RMSE is in real units.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{n}} \tag{5}$$

Equation 6 is used to calculate the Fourth Root Mean Quadrupled Error (R4MS4E). This metric records in real units the level of overall agreement between the observed and modelled datasets. It is a non-negative metric that has no upper bound and for a perfect model the result would be zero. It comprises a weighted measure of the error in which marked emphasis is placed upon the largest deviations between the observed and modelled values. R4MS4E is related to RMSE (Equation 5) in that higher order even powers can be used to further bias the evaluation process in favour of peaks and high(er) magnitude events, that will in most cases exhibit the greatest errors, and be insensitive to low(er) magnitude sequences (Blackie and Eeles, 1985: p 324). R4MS4E is also comparable to Mean Higher Order Error (MS4E; Abrahart and See, 2000) which is expressed in squared units whereas R4MS4E measurements are in real units. Cannas et al. (accepted) used this as one of several measures for evaluation of different forecasting models.

$$\mathrm{R4MS4E} = \sqrt[4]{\frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^4}{n}} \tag{6}$$

Equations 7 and 8 are used to calculate the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). AIC and BIC are model selection metrics in which some traditional evaluation measure is adjusted according to the number of free parameters in each model, p, and the number of data points that were used in its calibration, m. Each metric attempts to account for model complexities; a model, for example, with more parameters or more calibration data might be expected to forecast more accurately than a parsimonious model with fewer degrees of freedom (providing that it is calibrated correctly). Each metric attempts to find the minimal model that best explains the dataset, which can be contrasted with more traditional approaches to modeling, such as starting from a null hypothesis. Each metric quantifies the relative performance of a model on the understanding that a model with lots of parameters will provide a close fit to the dataset, but will have few degrees of freedom, and thus be of limited utility. This balanced approach discourages overfitting. Thus, a potential solution that contains lots of parameters, calibrated on a large amount of material, is penalised, while credit is given to simpler and less demanding models. AIC and BIC in this instance comprise modified and weighted versions of RMSE (Equation 5). The scores follow a similar format to RMSE; comprising non-negative metrics that have no upper bounds, the preferred model being the one that has the lowest value, with the result for a perfect model being zero. Neither of two evaluation measures will be in real units, and although each score attempts to record the level of overall agreement between the observed and modelled datasets, each metric nevertheless suffers form the same problems as RMSE; the result comprises a weighted measure of the error in which marked emphasis is placed upon the largest deviations between the observed and modelled values.

$$\mathrm{AIC} = m \ln(\mathrm{RMSE}) + 2p \quad (7)$$

$$\mathrm{BIC} = m \ln(\mathrm{RMSE}) + p \ln(m) \quad (8)$$

Equation 9 is used to calculate the number of sign changes (NSC; which is not to be confused with the Nash-Sutcliffe Coefficient that is sometimes abbreviated to 'NSC' - see Equation 18). This metric comprises a simple sequential count of the number of instances in which the sign of the residual changes throughout each series. It can be either positive or negative. It is unbounded and for a perfect model the result would be zero (although a score of zero does not necessarily imply a perfect model). It measures the level of overall agreement between the observed and modelled datasets in terms of "systematic error" and a "consistent outcome". The closer this value is to zero the more consistent the model is at either over-estimating or under-estimating the observed dataset. The maximum score is related to the size of the dataset, and would occur when each point in the series has been either over-estimated or under-estimated, in the opposite direction to that in which the previous data point in that series was either over-estimated or under-estimated i.e. the maximum score is *n*. However, a high score does not indicate a poor model in terms of accurate forecasts, since it measures crossover and not real differences in the magnitude of the error such that to criss-cross the observed record at each step by a small margin of error would in fact be a good thing to do in terms of overall fit. NSC is not related to the dimensions of the records that are being modelled. It is not weighted towards either high(er) magnitude or low(er) magnitude events and evaluates all deviations from the observed values in a sequential manner, based on sign, but is itself an *unsigned* measure of error. The principal use of this count is to increase the set of unrelated evaluation

metrics that are used to record the computed difference between simulated and observed datasets for model calibration purposes (Gupta *et al*., 1998).

$$\text{NSC} = \text{Number of Sign Changes (of residuals)} \tag{9}$$

2.2.2 Estimation of relative parameters

Equation 10 is used to calculate the Relative Absolute Error (RAE). This metric comprises the total absolute error made relative to what the total absolute error would have been if the forecast had simply been the mean of the observed values. It is a non-negative metric that has no upper bound and for a perfect model the result would be zero. It records as a ratio the level of overall agreement between the observed and modelled datasets. The lower the value the better the performance of the model compared to forecasting the mean of the series: a score of one indicates the same performance as forecasting the mean. This metric should be not considered in isolation since its evaluation of model output forecasts is related to the spread of the observed records. Thus, for a particular score, a model that is forecasting outputs within a limited range should be more accurate than a model that is forecasting across a broader range of values. RAE must not be confused with "Absolute Relative Error" (ARE; Hu *et al*., 2001) which does not perform a comparison with the mean of the time series.

$$\text{RAE} = \frac{\sum_{i=1}^{n} \left| Q_i - \hat{Q}_i \right|}{\sum_{i=1}^{n} \left| Q_i - \overline{Q} \right|} \tag{10}$$

Equation 11 is used to calculate the Percent Error in Peak (PEP). This metric comprises the difference between the highest value in the modelled dataset and the highest value in the observed dataset, made relative to the magnitude of the highest value in the observed dataset, and expressed as a percentage. It can be either positive or negative. It is unbounded and for a perfect model the result would be zero. PEP does not attempt to represent in a direct manner the level of overall agreement between the two datasets, and the temporal relationship that exists between the maximum value in each dataset is not considered. Like PDIFF, there is no requirement for the two peak values to coincide. PEP would thus appear to be more appropriate for single-event modelling as opposed to continuous modelling and care must be taken when applying this metric to continuous hydrographs since the two maximums that are being compared might derive from different storm events in the observed and modelled datasets The test could also be modified to perform a comparison that was limited to a consideration of annual maximums such that numerous computations would be performed on temporal subsets of the full record e.g. Sudheer *et al*. (2003). Equation 11 is in fact a modified version of the original simple "percent error in peak" that was recommended for the comparison of single-event models in the major review of evaluation criteria that was conducted by Green and Stephenson (1986) and thereafter endorsed by ASCE (1993). The metric in this case has been adjusted by swapping over the observed and modelled values in the numerator so that under-estimates produce a positive value and over-estimates produce a negative value i.e. equivalent to PDIFF (Equation 2). PEP differs from PDIFF in that the former has been divided by the maximum observed value.

$$\mathrm{PEP} = \frac{\max(Q_i) - \max(\hat{Q}_i)}{\max(Q_i)}.100 \ \ [\text{for } i = 1 \text{ to } n] \tag{11}$$

Equation 12 is used to calculate the Mean Absolute Relative Error (MARE). This metric comprises the mean of the absolute error made relative to the observed record. It has also been termed "Relative Mean Error" (RME; Khalil *et al*., 2001) and, in conflict with Equation 14, "Mean Relative Error" (Karunanithi *et al*., 1994; Elshorbagy *et al*., 2000; Teegavarapu and Elshorbagy, 2005). It is a non-negative metric that has no upper bound and for a perfect model the result would be zero. It records as a ratio the level of overall agreement between the observed and modelled datasets and is often expressed in percentage terms; the Mean Absolute Percent Error (MAPE; Armstrong and Collopy, 1992; Maier and Dandy, 1996; Bowden *et al.*, 2002). MARE is a relative metric which is sensitive to the forecasting errors that occur in the low(er) magnitudes of each dataset. In this case, because the errors are not squared, the evaluation metric is less sensitive to the larger errors that usually occur at higher magnitudes. It is nevertheless subject to potential 'fouling' by small numbers in the observed record. The principal difference between RAE (Equation 10) and MARE (Equation 12) is that the latter measurement is expressed in units that are relative to the observed record, as opposed to units that are relative to variation about the mean of the observed record, which could be difficult to interpret in an operational or decision-making context (Makridakis, 1993).

$$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Q_i - \hat{Q}_i\right|}{Q_i} \qquad (12)$$

Equation 13 is used to calculate the Median Absolute Percentage Error (MdAPE). This metric comprises the median of the absolute error made relative to the observed record. It is a non-negative metric that has no upper bound and for a perfect model the result would

be zero. It records as a ratio the level of overall agreement between the observed and modelled datasets and is similar to MARE (Equation 12). However, being based on the median, as opposed to the mean, this metric is less affected by skewed error distributions and the detrimental impact of problematic outliers. Trimming, in which high and low errors are discarded, in this case removes from consideration all values that are higher and lower than the middle ranked value. It thus helps to counter, in a radical manner, the powerful effects of high(er) magnitude errors and low(er) magnitude fouling whilst at the same time, serving to reduce the natural bias that exists in favour of low(er) output forecasts (Armstrong and Collopy, 1992; Fildes, 1992; Gardner, 1983). MdAPE also provides a standard trimming rule that can assist in the comparison of different studies (Armstrong and Collopy, 1992).

$$\text{MdAPE} = \text{Median}\left(\left|\frac{Q_i - \hat{Q}_i}{Q_i}\right| \text{x } 100\right) \tag{13}$$

Equation 14 is used to calculate the Mean Relative Error (MRE). This metric comprises the mean of the error made relative to the observed record. It is a signed metric that is unbounded and for a perfect model the result would be zero. It records as a ratio the level of overall agreement between the observed and modelled datasets. However, a low score does not necessarily indicate a good model in terms of accurate forecasts, since positive and negative errors will tend to cancel each other out and, for this reason, MARE (Equation 12) or MdAPE (Equation 13) are more popular metrics.

MRE is a relative metric which is sensitive to the forecasting errors that occur in the low(er) magnitudes of each dataset. In this case, because the errors are not squared, the

evaluation metric is less sensitive to the larger errors that usually occur at higher values. It is nevertheless subject to potential fouling by small numbers in the observed record.

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_i - \hat{Q}_i}{Q_i}\right) \tag{14}$$

Equation 15 is used to calculate the Mean Squared Relative Error (MSRE). This metric comprises the mean of the squared relative error in which relative error is error made relative to the observed record. It is a non-negative metric that has no upper bound and for a perfect model the result would be zero. It records as a ratio the level of overall agreement between the observed and modelled datasets. However, a low score does not necessarily indicate a good model in terms of accurate forecasts, since positive and negative errors will tend to cancel each other out and for this reason MARE (Equation 12) or MdAPE (Equation 13) are more popular metrics. MSRE works in a similar manner to MRE (Equation 14) in that this metric provides a relative measure of model performance, but in this case, the use of squared values makes it far more sensitive to the larger relative errors that will occur at low(er) magnitudes. It will in consequence be less critical of the larger absolute errors that tend to occur at higher magnitudes and more prone to potential fouling by small numbers in the observed record. The scale of the records that are to be analysed is thus important; moderate to low(er) values throughout (e.g. less than 100 cumecs) would make it more difficult for a model to achieve a superior score in comparison to one that was forecasting high(er) events (e.g. over 1000 cumecs). The high(er) magnitude model would have to make some particularly poor estimates to achieve a similar score to the low(er) magnitude model that could be predicting the actual

values to within 20 cumecs. It is a sensitive metric and special care should be taken in performing quantitative comparisons between different studies and different catchments.

$$\mathrm{MSRE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_i - \hat{Q}_i}{Q_i}\right)^2 \qquad (15)$$

Equation 16 is used to calculate the Relative Volume Error (RVE). This signed metric comprises the total error made relative to the total observed record. It is unbounded and for a perfect model the result would be zero. It records as a ratio the level of overall agreement between the observed and modelled datasets and is a popular metric that is often expressed in percentage terms using different phrasing e.g. "Error in Volume" (Rajurkar *et al*., 2004); “Error of Total Runoff Volume” (EV; Lin and Chen, 2004); "Percent Bias" (PBIAS; Yapo *et al*., 1996; Yu and Yang, 2000); "Deviation of Runoff Volumes" (Dv; WMO, 1986; ASCE, 1993). RVE is a relative measure of overall volume error that is used to provide an indication of the overall water balance of the model – i.e. it is equivalent to the difference in mean flow over the period. RVE has similar properties to MRE and MSRE in that a low score does not necessarily indicate a good model, in terms of accurate forecasts, since positive and negative errors will tend to cancel each other out. Indeed, a perfect model will return a score of zero, but it is also possible for a model that bears no resemblance to the actual hydrograph to produce a zero.

RVE was considered to be an "adequate measure" for the volumetric assessment of single event models (Green and Stephenson, 1986); but, conversely, it is a recommended metric for the evaluation of continuous hydrographs (ASCE, 1993). RVE must not be confused

with similar sounding metrics such as 'Percent Error in Volume' (PEV; Paik *et al*., 2005) which is an unsigned equivalent.

$$\mathrm{RVE} = \frac{\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)}{\sum_{i=1}^{n} Q_i} \tag{16}$$

2.3 Dimensionless coefficients that contrast model performance with accepted norms or standards

Equation 17 is used to calculate the 'Coefficient of Determination' which is the square of the 'Pearson Product Moment Correlation Coefficient' (RSqr; Pearson, 1896). This metric comprises the squared ratio of the combined dispersion of two series to the total dispersion of the observed and modelled series. It describes the proportion of the total statistical variance in the observed dataset that can be explained by the model. It ranges from 0.0 (poor model) to 1.0 (perfect model). It records as a ratio the level of overall agreement between the observed and modelled datasets; the equation, however, is based on a consideration of linear relationships and limited in that it standardizes to the observed and modelled means and variances. The metric is insensitive to additive and proportional differences between the observed and modelled datasets, such that high scores can be obtained, even if the simulated values are considerably different from the observed values in terms of magnitude and variability. Indeed, since quantification is restricted to a consideration of differences in dispersion, a solution with systematic errors that over-estimated or under-estimated on each occasion would still produce a good result even if all of the numbers were wrong. The model in such cases would exhibit serious

flaws that should, but does not, preclude it from being assigned a "near perfect" score. The metric is also oversensitive to outliers and thus biased towards a consideration of extreme events such that the true overall relationship is obscured. The limitations of this metric and other correlation-based measures are well documented (e.g. Kessler and Neas, 1994; Legates and Davis, 1997; Legates and McCabe, 1999); it was, nevertheless, still 'common practise' to use such measures in the 1990s (Chiew and McMahon, 1993). To redress such quandaries it is possible to make better use of additional material such as the intercept and gradient of the regression equation upon which this metric is based. For good agreement, the intercept should be close to zero, and the gradient can be used to provide a weighted version of this metric (wRSqr; Krause *et al*., 2005). This metric is a basic statistical method and its output can be tested for 'statistical significance'. Testing would involve the use of traditional parametric procedures and requirements not least of which are the assumptions of a bivariate normal distribution and a homoscedastic relationship.

$$\mathrm{RSqr} = \left[\frac{\sum_{i=1}^{n} (Q_i - \overline{Q})(\hat{Q} - \widetilde{Q})}{\sqrt{\sum_{i=1}^{n} (Q_i - \overline{Q})^2 \sum_{i=1}^{n} (\hat{Q}_i - \widetilde{Q})^2}} \right]^2 \qquad (17)$$

Equation 18 is used to calculate the Coefficient of Efficiency (CE; Nash and Sutcliffe, 1970). This popular metric has several aliases and is sometimes confused with other metrics: *Determination Coefficient* or *Coefficient of Determination*, the *Efficiency Index*, the *F index* or the *Nash-Sutcliffe Coefficient* R^2. It is one minus the ratio of Sum Square Error (SSE) to the statistical variance of the observed dataset about the mean of the observed dataset. CE is intended to range from zero to one but negative scores are also

permitted. The maximum positive score of one represents a perfect model; a value of zero indicates that the model is no better than a one parameter 'no knowledge' model in which the forecast is the mean of the observed series at all time steps; negative scores are unbounded and a negative value indicates that the model is performing worse than a 'no knowledge' model. If the last observed record was an input to the model a negative score would suggest that the model is degrading the information that has been provided. It records as a ratio the level of overall agreement between the observed and modelled datasets and represents an improvement over the Coefficient of Determination (RSqr) for model evaluation purposes since it is sensitive to differences in the observed and modelled means and variances. For such reasons it carries strong past support (e.g. NERC, 1975; ASCE, 1993). It nevertheless suffers from certain similar handicaps, owing to the use of squared differences that are sensitive to peak flow values and insensitive to low(er) magnitude conditions. CE is also insensitive to systematic positive or negative errors and has been criticised for interpretational difficulties, since even poor models can produce relatively high values and the best models do not produce values that on first examination are that much higher (Garrick *et al*., 1978: p376; Krause *et al*., 2005). It will also produce optimistic results in cases where the hydrological regime of interest exhibits marked seasonal variations, such that intrinsic periodic variation is an important part of total observed variation (Garrick *et al*., 1978; Lorrai and Sechi, 1995). The Modified Coefficient of Efficiency (MCE; Legates and McCabe, 1999; Krause *et al*., 2005) is a more generic metric that could be of particular interest since, through the use of absolute values, it will permit errors and differences to be given a more appropriate weighting i.e. no inflated squared values. It is also possible to construct a relative Coefficient of Efficiency (E_{rel}; Krause *et al*., 2005). Beran (1999) and Garrick *et al*. (1978: p376)

present strong arguments against using the observed mean to develop a horizontal hydrograph for modelling comparison purposes; better methods exist to define the baseline against which a model should be compared e.g. use of persistence or seasonal averages. No simple equation exists to determine the statistical significance of this metric and bootstrap methods are required to address such matters: see Efron (1981a, b), Efron and Gong (1983), or Willmott *et al*. (1985).

$$\mathrm{CE} = 1 - \frac{\sum_{i=1}^{n} (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n} (Q_i - \overline{Q})^2} \qquad (18)$$

Equation 19 is used to calculate the Index of Agreement (IoAd; Wilmott, 1981). This metric is one minus the ratio of Sum Square Error (SSE) to "potential error" in which potential error represents the sum of the 'largest quantification' that can be obtained for each individual forecast with respect to the mean of the observed dataset. It is based on a two part squared distance measurement in which the absolute difference between the model-simulated value and the observed mean is added to the absolute difference between the observed record and the observed mean. The result is then squared and summed over the series. IoAd ranges from 0.0 (poor model) to 1.0 (perfect model). It records as a ratio the level of overall agreement between the observed and modelled datasets and also represents an improvement over the Coefficient of Determination (RSqr) for model evaluation purposes since it is sensitive to differences in the observed and modelled means and variances. Its outputs are similar in range and interpretation to the Coefficient of Determination (RSqr); but dissimilar in range and interpretation to the Coefficient of Efficiency, since despite being based on the mean, IoAd has no meaningful

zero to provide a convenient reference point against which to compare model outputs in terms of forecasting capabilities with respect to the observed mean. It was a purposeful attempt to overcome the shortcomings of past metrics that were considered to be insensitive to differences in the observed and model-simulated means and variances whilst at the same time attempting to retain the use of fixed upper and lower bounds and thus avoid the problem of negative scores and subjective interpretations of "performing worse than a 'no knowledge' model". It also suffers from certain similar handicaps, owing to the use of squared differences that are sensitive to peak flow values and insensitive to low(er) magnitude conditions. IoAd is also insensitive to systematic positive or negative errors and has been criticised for interpretational difficulties, since even poor models can produce relatively high values, and the best models do not produce values that on first examination are that much higher (Willmott *et al*., 1985; Krause *et al*., 2005). The Modified Index of Agreement (MIoAd; Willmott *et al*., 1985; Legates and McCabe, 1999; Krause *et al*., 2005) is a more generic metric that could be of particular interest since through the use of absolute values it will permit errors and differences to be given a more appropriate weighting i.e. no inflated squared values. It is also possible to construct a relative Index of Agreement (d_{rel}; Krause *et al*., 2005).

$$\text{IoAd} = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(\left|\hat{Q}_i - \overline{Q}\right| + \left|Q_i - \overline{Q}\right|)^2} \qquad (19)$$

Equation 20 is used to calculate the Coefficient of Persistence (Kitanidis and Bras, 1980). The use of this metric is on the increase and it is often referred to in published papers under different aliases such as the Persistence Index (PI) or G_{bench}. It is one minus the

ratio of Sum Square Error (SSE) to what Sum Square Error (SSE) would have been if the forecast had been the last observed value. It has strong similarities with CE (Equation 18); in this instance, however, the last observed record is used instead of the observed mean for the purposes of model comparison. PI is intended to range from zero to one but negative scores are also permitted. The maximum positive score of one represents a perfect model; a value of zero indicates that the model is no better than a one parameter 'no knowledge' model in which the forecast is the last observed record at all time steps and equates to 'a no change situation'; negative scores are unbounded and a negative value indicates that the model is performing worse than a 'no knowledge' model. This metric compares model outputs with forecasts obtained by assuming that the process being modelled is a Wiener process i.e. variance increases linearly with time and so the best estimate for the future is given by the latest measurement. It records as a ratio the level of overall agreement between the observed and modelled datasets and represents an improvement over both the Coefficient of Efficiency and the Coefficient of Agreement. For lead time model forecasting assessments a more meaningful value, or floor against which to measure poor levels of fit, is required since testing relationships in terms of variations about the mean of the observed series is neither stringent nor rigorous and produces results that are difficult to interpret (Anctil *et al.*, 2004). This metric suffers from the problem of negative scores and subjective interpretations of “performing worse than a 'no knowledge' model”. It also suffers from certain similar handicaps, owing to the use of squared differences that are sensitive to peak flow values and insensitive to low(er) magnitude conditions.

PI can be interpreted as a special case with respect to the 'benchmark proposals' of Seibert (2001). It was suggested that the performance of a model should be compared with the performance of a target model, as opposed to a constant mean. The target model could be anything devised by the user - such as the observed runoff shifted backwards by one or more time steps - that, at the first time step, equates to this particular metric. The same comments apply regarding statistical significance as with CE and IoAd.

$$\mathrm{PI} = 1 - \frac{\sum_{i=1}^{n} (Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n} (Q_i - Q_{i-1})^2} \tag{20}$$

2.4 Summary

Table 2 summarizes some properties of the statistical measures introduced above. It provides the 'best' possible value for the statistic (i.e. produced when applying the measure to a perfect model) and the 'worst' possible theoretical value (quite often $\pm\infty$ but, in practice, models are unlikely to result in a value anywhere near this magnitude). This table also presents some subjective indicators of values that might represent 'good' models, 'satisfactory' models and 'poor' models. This use of non-specific nomenclature is similar to the approach of Shamseldin (1997) who considered models with a CE score of 0.9 and above to be 'very satisfactory', between 0.8 to 0.9 to be 'fairly good', and below 0.8 to be 'unsatisfactory'.

In cases where the statistic is data dependent (i.e. its magnitude varies according to the scale or quantity of data that are analysed) this value cannot be determined and is represented by DD. In these cases it would be wrong to draw fixed conclusions from the

calculated statistic which should not be used to perform cross catchment comparisons between different studies. However, data dependent statistics can provide measures that can be used to compare models based on the same catchment and dataset.

Note that although *Good*, *Satisfactory*, and *Poor* figures are given for RSqr these values are actually dependent on the number of data points analysed. The statistical significance of the RSqr relationship should be calculated separately – these figures merely provide a guide for large datasets.

3. Examples of model error

Given the wide array of performance metrics, a problem facing the hydrological modeller is deciding which error measure or measures are most appropriate for a particular type of hydrological application. Hall (2001), for example, stressed the importance of using a set of measures that was focussed on important aspects of model behaviour; whereas ASCE (1993) reviewed various measures of fit with respect to hydrological models and recommended the use of three metrics for continuous hydrograph time series flow predictions and four metrics for single-event hydrographs in which the response of a catchment to a single storm is predicted. Chiew and McMahon (1993) showed that high levels of overlap can occur in solutions that are considered to be "perfect"; CE scores $\geq$0.93 corresponded to RSqr values $\geq$0.97 or RSqr values $\geq$0.93 in cases where mean simulated flow was within 10% of mean recorded flow. It is also possible to develop a purposeful integration of different evaluation metrics: Tang *et al*. (2006) combined weighted Root Mean Squared Error (RMSE) values computed on different partitions of the discharge record (low, average, high) for use in the calibration process; Mean Squared

Error (MSE) has been combined with Mean Absolute Relative Error (MARE; see previous comments on MRE) to produce Pooled Mean Squared Error (PMSE; Elshorbagy *et al*., 2000) and Fuzzy Mean Squared Error (FMSE; Teegavarapu and Elshorbagy, 2005) for use in the testing process. Watts (1997: p176) also noted that 'in a flood routing model accurate prediction of time to peak flow may be essential'; whereas in a model used for water resource allocation 'long-term low flows may be of interest'. It is thus important that modellers select appropriate and relevant evaluation measures from the set of existing metrics that were discussed earlier. The modeller must be placed in an informed position, such that the decision to select a specific metric is based on the particular requirements of each individual application. The modeller will also on occasion need to resolve conflicting evaluations.

Figure 1 (adapted from Dawson and Wilby, 2001) helps to highlight the dilemma facing the modeller by showing different types of model error related to four sets of hypothetical flow forecasting model outputs. The hypothetical model output datasets are available for download from the HydroTest web site in text format. It is hoped that other interested parties might develop improved evaluation metrics based on these four hypothetical datasets, or submit a recorded time series dataset for inclusion on the web site as a 'benchmark series', such that the international research effort will in due course use such donated datasets or other relevant international datasets to develop a set of cross-catchment standards for model testing and the quantification of model forecasting skill (Seibert, 2001).

Hypothetical Model A, which is somewhat naive, predicts the shape of the hydrograph well but consistently over-estimates the flow and appears to model the observed flow late – the *Naïve Model*. Hypothetical Model B follows low flow regimes accurately but misses the flood peak completely – the *Low Flow Model*. Hypothetical Model C follows the hydrograph reasonably closely, but contains a lot of noise; either over- or under-predicting each flow event by around 15% – the *Noisy Model*. Hypothetical Model D, although capturing the flood peak well, is weak at modelling low flows – consistently over-estimating low flow events – the *High Flow Model*. Presented with these four hypothetical models, the dilemma facing the hydrologist is in deciding upon which, if any, measure(s) will provide an objective indication of the most appropriate model to use under different hydrological circumstances or with which to perform specific modelling operations. From a visual inspection of the hydrographs it is relatively easy to categorise the four different types of model, but scientists require more objective measures of model performance – especially when the different types of model output error are less distinctive than in these simple illustrative examples.

Table 3 presents the error measures discussed in the previous section for the four hypothetical models. Those values highlighted in bold in this table indicate the 'best' model out of the four when assessed using each particular evaluation metric. It can be seen from this table that no one model is consistently 'best' in terms of the numerous evaluation metrics, although some models appear to be 'better' than others, and various trade-offs exist.

Take, as an example, two popular evaluation metrics – Mean Squared Relative Error (MSRE) and Root Mean Squared Error (RMSE). MSRE measures relative performance and is thus more critical of errors that occur at low flows. For example, a model that predicts a flow of 2 cumecs when the observed flow is 4 cumecs is more heavily penalised than a model that predicts a flow of 102 cumecs when the observed flow is 104 cumecs. Hypothetical Model B appears in Figure 1 to model the low flow events most accurately and, unsurprisingly, has the lowest MSRE value of 0.0208. It nevertheless misses the flood peak by a considerable margin. The Naïve Model (A), conversely, generally over estimates all of the observed values and has the 'worst' MSRE (0.0510).

RMSE, on the other hand, measures overall performance across the entire range of the dataset. It is sensitive to small differences in model performance and being a squared measure exhibits marked sensitivities to the larger errors that occur at higher magnitudes. In this case, because Hypothetical Model C generally follows the hydrograph across the full range of flow events, it has the smallest RMSE. This simple example illustrates the dangers of relying on one measure alone to evaluate and select between different models. Scientists should use a range of methods to evaluate their models, including subjective visual inspection of the forecast outputs, in order to inspect the relevant features of the functional behaviour in which they are interested.

Looking at each model in turn, Hypothetical Model A – the Naïve Model – is not identified as the 'best' model by any of the evaluation metrics. However, there are times when this model appears to perform well according to certain measures. For example, in terms of RSqr (which equates to 82.73%), the Naïve Model is the second 'most accurate'

of the four models. This echoes the results of Legates and McCabe (1999) who point out the fallibility of this evaluation metric which does not penalise additive and proportional differences.

Hypothetical model B – the Low Flow Model – scores well in terms of relative evaluation metrics such as MSRE, MAE and MARE but less well when assessed using indicators that are critical of models that are poor in forecasting high flow events; AME and PDIFF. Because this model misses the flood peak by some margin it is heavily penalised by a number of other error measures too (i.e. it is ranked least accurate); RMSE, CE, R4MS4E and RSqr. These evaluation measures on their own might lead the modeller to discard such a model even though it is far more accurate than any other hypothetical model during the low flow periods.

Hypothetical Model C – the Noisy Model – appears to model the general shape of the hydrograph well but any individual forecast is only accurate to within around 15% of the observed record. Consequently, this model performs well according to RMSE, CE, ME, and PI, but badly when evaluated using NSC.

Hypothetical Model D – the High Flow Model – forecasts the peak flow accurately and, although over-estimating the low flow events, follows the low flow period closely. As one would expect this High Flow Model scores well for PDIFF and AME, which focus on the assessment of the peaks, and badly with MdAPE. This model also scores well in terms of RSqr – primarily because it follows the general shape of the hydrograph well.

It is worth noting that all four sets of example model outputs have returned a negative PI value, indicating that the models appear to be 'degrading' the input dataset. Clearly, if the models were attempting to make one-step-ahead forecasts the modeller would be advised to use a Naïve one-step-ahead model based on this index. However, in the example datasets, the lead time for the models has not been specified and care must taken when evaluating a $t+n$ model using this index when n is greater than 1. Ideally the PI should be adapted so that the model under evaluation is compared with the equivalent *feasible* naïve time-delayed model (i.e. Q_{i-1} in the denominator of Equation 20 is changed to Q_{i-t} where t is the lead time of the model under investigation) or with an appropriate benchmark series such as those proposed by Seibert (2001).

4. Case study

In this section the application of the HydroTest web site to the real problem of developing and evaluating a rainfall-runoff model for the River Ouse in northern England is discussed. The model developed for this catchment was an Artificial Neural Network (ANN) which required a number of parameters to be established during calibration. While the detail of this kind of model is beyond the intended scope of this paper, a brief summary of its calibration process is presented here.

The usual approach to developing and evaluating such models involves the use of three datasets – one for training a series of different models (the *training* dataset), one for validating these models and selecting the 'best' one (the *validation* dataset), and one for the final 'unseen' testing of the chosen model (the *test* dataset). During calibration the modeller is attempting to find the 'best' configuration for the ANN model (how many

hidden nodes it should have in its structure) and also the optimum training period (ANNs are trained by exposing them to the training dataset a number of times – called *epochs*). Thus, the modeller creates a number of different models (optimised according to the mean squared error) – with different structures, trained for different periods using the training dataset - before selecting the 'best' model by evaluating all these models against the validation dataset.

The data for this study come from the River Ouse which is a catchment of 3315km^2 containing an assorted mix of urban and rural land uses. It exhibits a significant amount of natural variation, from dissected uplands in the west that experience substantial precipitation, to cultivated lowlands in the east with more equitable weather. The City of York is the main point of concentrated development in the catchment (population of 181,094; 27,200 hectares). Hourly data were available for this study from three winter periods (spanning 1 October to 31 March) and were split according to the study of Dawson *et al*. (2006b) as follows; 1993/94 was used for training; 1995/96 was used for validation; 1994/95 was used for testing. Flow data were available for three upstream sites at Crakehill, Skip Bridge and Westwick along with data from five rainfall gauges at Tow Hill, Arkengartdale, East Cowton, Osmotherly and Malham Tarn. These data were analysed and processed to provide nine predictors of flow (measured in cumecs) for the River Ouse at Skelton with a lead time of 24 hours.

During training a number of different networks were developed. They were trained from between 200 to 4000 epochs in steps of 200 epochs (a range known to be effective based on past experience), using 2, 3, 4 and 5 hidden nodes (also chosen based on past

experience). The subsequent trained models were then analysed and the validation dataset used to identify the 'best' model based on various statistics that are available on the HydroTest web site.

Figure 2 illustrates the problem encountered by the modeller when attempting to decide on the 'best' model for implementation. In this figure the values of MSRE, RMSE, CE and MAE are presented for each neural network configuration (2, 3, 4, 5 hidden nodes) and for each training period (200 – 4000 epochs). Although it appears that the four node model appears to consistently produce the 'best' results according to these four metrics, there is still some uncertainty as to how long the model should be trained for. Figure 2 highlights the optimum result for each error measure – identified by the arrow. RMSE and CE statistics identify the 'best' model as that trained for only 400 epochs, MSRE identifies the best model as that trained for 1400 epochs, while MAE identifies the model trained for 1200 epochs as the most accurate when assessed with the validation dataset.

The modeller must now decide which of these statistics gives her/him the best indication of model performance for the chosen application. The modeller may be tempted by the 400 epoch model as this appears to be the most accurate according to two evaluation metrics. The 1400 epoch model is perhaps the 'best' model for modelling low flows according to the MSRE. However, the 1200 epoch model appears to provide the best absolute model. A compromise could involve the selection of a model trained for say 800 or 1000 epochs as this lies around midway between the three models (400, 1200, 1400 epoch models).

In fairness there is little to choose between any of these models, since the measures are very close for all identified configurations, and selection of any one of them would make little difference in a practical application. However, the example does show that the metrics do measure different characteristics of the model and the user should be aware of the strengths and limitations of each option. The metrics should not be applied 'blindly' without some form of deeper understanding of their application and interpretation. It would be wrong to focus on any one metric and, as Jakeman et al. (2006) note; a wide range of performance indicators should be examined.

5. The HydroTest Web Site

The HydroTest web site is available at www.hydrotest.org.uk. Following an initial registration page, the user is directed towards the site's home page which introduces the site and provides links to the five main areas – *Details* (which provides the user with information about the statistics and how the site operates), *References* (a list of appropriate references), *Benchmarks* (discussion on the four example outputs and download access to the example datasets in text format), *Analysis* (the analysis pages themselves) and a discussion forum. In this section the *Analysis* pages and process are discussed.

The first step of the 'Analysis Process' (Figure 3) involves the user selecting and uploading their data for analysis (note that the system only stores these data temporarily – they are deleted from the system once the analysis is complete). The user should have two sets of data for processing – a set of observed (i.e. actual) data and a set of modelled (i.e. forecast or predicted) data. On the user's system these data can be stored in either

one text file – tab delimited or comma separated (observed data in the first column, modelled in the second column) – or in two separate text files (one file containing a single column of observed data and the other file containing a single column of modelled data). The files should not contain any non-numeric values; for example, column titles at the head of the text files, as these will invalidate the calculations.

The user uses the 'Browse' buttons to select the file(s) on their own computer and, if the data are contained in a single file, the user selects the appropriate radio button to indicate how the data are delimited. The web site will assume that if only one data file is selected (which must be in the observed input box) that this file will contain both sets of data in two columns. The user then selects 'Upload' to load the data into the system for analysis. The *Analysis Step 2* screen is then displayed (Figure 4).

Step 2 of the process allows the user to set a number of parameters before the analysis takes place (although the default values set on this page can be left as they are). First, if the dataset contains any missing values represented by a specific code number, this value should be entered here (the default is -999). Secondly, the user should select how many decimal places the results will be displayed to (the default is four).

If the user wishes the statistics to be calculated on a subset of the data (i.e. only assessing data that fall within a certain range) they should click the checkbox and enter the lower and upper bounds in the boxes below. This will limit the analysis such that metrics are only calculated, and the two datasets are only compared, when the observed data point falls within this boundary. The modelled data point can be outside this range and the

calculation will still be performed provided the observed value is acceptable (i.e. not missing). Values must be strictly less than the lower bound and strictly greater than the upper bound for the data point to be excluded from the calculation (i.e. the data point will be included in the analysis if it is equal to the lower or upper bound).

The final two fields (free parameters and data points) should be completed if the user wishes the AIC and BIC statistics to be calculated (the user is requested to enter the number of free parameters in the model and the number of data points used in calibrating the model – both are required for calculating AIC and BIC). Clicking 'Calculate' performs the analysis and leads to the final results screen (Figure 4).

The results are presented in the format shown in Figure 4 (note, part of this screen has been scrolled out of view so that some of the results can be seen). This screen shows the total number of data points the system analysed in the files (the number of rows read) and, for reference, the missing code identifier and the name(s) of the file(s) analysed. The initial set of output statistics provides descriptive measures of the observed dataset (mean, maximum, variance, etc.) and the number of missing values encountered in the file. If the user selected a range of data for analysis, the number of observed data points encountered outside this range is also noted. These statistics are then followed by the equivalent statistics for the modelled dataset.

The following section of the screen presents all the calculated statistics discussed in Section 2. Finally, the user can download a copy of these results as a text file by clicking on the appropriate link at the bottom of the page.

6. Conclusions

This paper has introduced the HydroTest web site – a site which provides a consistent collection of evaluation metrics and example datasets for modellers to use in subsequent assessment exercises. The discussion has highlighted the importance of not relying on individual measures of performance to evaluate data series. It is suggested that hydrological modellers should report a minimum set of CE, RSqr and IoAd statistics in addition to the provision of other more specific data dependent measures (such as RMSE, MSR, etc.). This will facilitate subsequent comparisons to be made between different catchments and studies. Future developments for the site include the implementation of a data plotting facility for time series and scatter diagrams; the added functionality of analysing multiple files simultaneously; the continued development and implementation of new statistics (such as variations to existing metrics discussed in this paper); and the monitoring of the site's performance and its promotion as a standardised evaluation toolbox.

There remains an open invitation to users to;

- Provide any comments and improvements / corrections to the site;
- Identify any missing statistics that could be incorporated into the site;
- Develop and submit additional hydrological error measures that may enable more suitable cross-study comparisons than exist already;

- Develop and submit additional hydrological error measures for the example datasets that are provided on the web site.
- Develop and submit further hydrological datasets for inclusion on the web site as a 'benchmark series'.

7. References

Abrahart, R.J. and See, L.M. 2000. 'Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments', Hydrological Processes, Vol 14(11-12), pp 2157 – 2172.

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

Alvisi, S., Mascellani, G., Franchini, M. and Bardossy, A. 2006. 'Water level forecasting through fuzzy logic and artificial neural network approaches', Hydrology and Earth System Sciences, Vol 10(1), pp 1 – 17.

Anctil, F., Michel, C., Perrin, C. and Vazken Andréassian, V. 2004. 'A soil moisture index as an auxiliary ANN input for stream flow forecasting', Journal of Hydrology, Vol 286(1-4), pp 155 – 167.

Anderson, M.G. and Burt, T.P. 1985. Hydrological Forecasting. John Wiley & Sons Ltd. - Chichester, UK.

Armstrong, J.S. and Collopy, F. 1992. 'Error measures for generalizing about forecasting methods: Empirical comparisons', International Journal of Forecasting, Vol 8(1), pp 69 – 80.

Atiya, A.F., El-Shoura, S.M., Shaheen, S.I. and El-Sherif, M.S. 1999. 'A Comparison Between Neural-Network Forecasting Techniques - Case Study: River Flow Forecasting', IEEE Transactions on Neural Networks, Vol 10(2), pp 402 – 409.

ASCE. 1993. 'Criteria for Evaluation of Watershed Models', Journal of Irrigation and Drainage Engineering, Vol 119(3), pp 429 – 442.

ASCE. 2000. 'Artificial Neural Networks in Hydrology. II: Hydrologic Applications', Journal of Hydrological Engineering, Vol 5(2), pp 124 – 137.

Baratti, R. Cannas, B. Fanni, A. Pintus, M. Sechi, G.M. and Toreno, N. 2003. 'River flow forecast for reservoir management for neural networks', Neurocomputing, Vol 55, pp 421 – 437.

Beran, M. 1999. 'Hydrograph Prediction – How much skill?', Hydrology and Earth System Sciences, Vol 3(2), pp 305 – 307.

Blackie, J.R. and Eeles, C.W.O. 1985. 'Lumped catchment models' , in Hydrological Forecasting, Anderson, M.G. and Burt, T.P. (eds), Wiley, UK. pp 311 – 345.

Boughton, W. 2006. 'Calibrations of a daily rainfall-runoff model with poor quality data', Environmental Modelling and Software, Vol 21, pp 1114 – 1128.

Bowden G.J., Maier, H.R. and Dandy, G.C. 2002. 'Optimal division of data for neural network models in water resources applications', Water Resources Research, Vol 38(2), 1010, doi:10.1029/2001WR000266.

Campolo, M., Soldati, A. and Andreussi, P. 2003. 'Artificial neural network approach to flood forecasting in the River Arno', Hydrological Sciences Journal, Vol 48(3), pp 381 – 398.

Cannas, B., Fanni, A., See, L.M. and Sias, G. accepted.. The role of data preprocessing for river flow forecasting using neural networks. Physics and Chemistry of the Earth.

Chang, F-J., Hu, H-F. and Chen, Y-C. 2001. 'Counterpropagation fuzzy-neural network for streamflow reconstruction', Hydrological Processes, Vol 15(2), pp 219 – 232.

Chang, F-J, Chang, L-C. and Huang, H-L. 2002. 'Real-time recurrent learning neural network for stream-flow forecasting', Hydrological Processes, Vol 16(13), pp 2577 – 2588.

Chang, L-C., Chang, F-J. and Chiang, Y-M. 2004. ‘A two-step-ahead recurrent neural network for stream-flow forecasting’, Hydrological Processes, Vol 18(1), pp 81 – 92.

Chen, S-H., Lin, Y-H., Chang, L-C. and Chang, F-J. 2006. ‘The strategy of building a flood forecast model by neuro-fuzzy network’, Hydrological Processes, Vol 20(7), pp 1525 – 1540.

Chibanga, R., Berlamont, J. and Vandewalle, J. 2003. ‘Modelling and forecasting of hydrological variables using artificial neural networks: the Kafue River sub-basin’, Hydrological Sciences Journal, Vol 48(3), pp 363 – 379.

Chiew, F.H.S. and McMahon, T.A. 1993. ‘Assessing the Adequacy of Catchment Streamflow Yield Estimates’, Australian Journal of Soil Research, Vol 31(5), pp 665 – 680.

Clarke, R.T. 1973. ‘A review of some mathematical models used in hydrology, with observations on their calibration and use’, Journal of Hydrology, Vol 19(1), pp 1 – 20.

Clements, M.P. and Hendry, D.F. 1993. ‘On the Limitations of Comparing Mean Square Forecast Errors’, Journal of Forecasting, Vol 12(8), pp 617 – 637.

Corani, G. and Guariso, G. 2005. ‘An application of pruning in the design of neural networks for real time flood forecasting’, Neural Computing and Applications, Vol 14(1), pp 66 – 77.

Coulibaly, P. Anctil, F. and Bobée, B. 2000. 'Daily reservoir inflow forecasting using artificial neural networks with stopped training approach', Journal of Hydrology, Vol 230(3-4), pp 244 - 257.

Dawson, C.W. and Wilby, R.L. 2001. ‘Hydrological modelling using artificial neural networks’, Progress in Physical Geography, Vol 25(1), pp 80 – 108.

Dawson, C.W. Harpham, C. Wilby, R.L. and Chen, Y. 2002. 'An Evaluation of Artificial Neural Network Techniques for Flow Forecasting in the River Yangtze, China', Hydrology and Earth System Sciences, Vol 6(4), pp 619-626.

Dawson, C.W. Abrahart, R.J. Shamseldin, A.Y. and Wilby, R.L. 2006a. 'Flood estimation at ungauged sites using artificial neural networks', Journal of Hydrology, Vol 319, pp 391 – 409.

Dawson, C.W. See, L.M., Abrahart, R.J. and Heppenstall, A.J. 2006b. 'Symbiotic adaptive neuro-evolution applied to rainfall-runoff modelling in northern England', Neural Networks, Vol 19(2), pp 236 – 247.

DeVos, N.J. and Rientjes, T.H.M. 2005. 'Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation', Hydrology and Earth System Sciences, Vol 9(1-2), pp 111 – 126.

Diskin, M.H. and Simon, E. 1977. 'A procedure for the selection of objective functions for hydrologic simulation models', Journal of Hydrology, 34(1-2), 129 – 149.

Efron, B., 1981a. 'Nonparametric estimates of standard error: The jackknife, the bootstrap, and other methods', *Biometrika*, Vol 68(3), pp 589 –599.

Efron, B., 1981b. **'**Nonparametric standard errors and confidence intervals**',** *Canadian Journal of Statistics*, Vol 9(2), pp 139 **–** 172.

Efron, B., and Gong, G. 1983. **'**A leisurely look at the bootstrap, the jackknife and cross-validation**'**. *American Statistician*, Vol 37(1), pp 36 **–** 48.

Elshorbagy, A. Simonovic, S.P. and Panu, U.S. 2000. 'Performance Evaluation of Neural Networks for Runoff Prediction', Journal of Hydrologic Engineering, Vol 5(4), pp 424 – 427.

Fernando, D.A.K. and Jayawardena, A.W. 1998. 'Runoff Forecasting Using RBF Networks with OLS Algorithm', Journal of Hydrologic Engineering, Vol 3(3), pp. 203-209.

Fildes, R. 1992. 'The evaluation of extrapolative forecasting methods', International Journal of Forecasting, Vol 8(1), pp 81 – 98.

Furundzic, D. 1998. 'Application example of neural networks for time series analysis: Rainfall-runoff modeling', Signal processing, Vol 64(3), pp 383 – 396.

Gardner, E.S. 1983. 'The Trade-offs in Choosing a Time Series Method', Journal of Forecasting, Vol 2(3), pp 263 – 267.

Garrick, M., Cunnane, C. and Nash, J.E. 1978. 'A criterion of efficiency for rainfall-runoff models', Journal of Hydrology, Vol 36(3-4), pp 375 – 381.

Gaume, E. and Gosset, R. 2003. 'Over-parameterisation, a major obstacle to the use of artificial neural networks in hydrology? ', Hydrology and Earth System Sciences, Vol 7(5), pp 693 – 706.

Giustolisi, O. and Laucelli, D. 2005. 'Improving generalization of artificial neural networks in rainfall-runoff modelling', Hydrological Sciences Journal, Vol 50(3), pp 439 – 457.

Green, I.R.A. and Stephenson, D. 1986. 'Criteria for comparison of single event models', Hydrological Sciences Journal, Vol 31(3), pp 395 – 411.

Gupta, H.V., Sorooshian, S. and Yapo, P.O. 1998. 'Toward improved calibration of hydrological models: Multiple and noncommensurable measures of information', Water Resources Research, Vol 34(4), pp 751 – 763.

Hall, M.J. 2001. ‘How well does your model fit the data?', Journal of Hydroinformatics, Vol 3(1), pp 49 – 55.

Houghton-Carr, H.A. 1999. ‘Assessment criteria for simple conceptual daily rainfall-runoff models’, Hydrological Sciences Journal, Vol 44(2), pp 237 – 261.

Hsu, K-L., Gupta, H.V. and Sorooshian, S. 1995. ‘Artificial neural network modeling of the rainfall-runoff process’, Water Resources Research, Vol 31(10), pp 2517 – 2530.

Hu, T.S., Lam, K.C. and Ng, S.T. 2001 ‘River flow time series prediction with a range-dependent neural network’, Hydrological Sciences Journal, Vol 46(5), pp 729 – 745.

Jain, A. and Srinivasulu, S. 2004. ‘Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques’, Water Resources Research, Vol 40(4), W04302.

Jain, A. and Srinivasulu, S. 2005. ‘Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques’, Journal of Hydrology, Vol 317(3-4), pp 291 – 306.

Jakeman, A.J. Letcher, R.A. and Norton, J.P. 2006. ‘Ten iterative steps in the development and evaluation of environmental models’, Environmental Modelling and Software, Vol 21, pp 602 – 614.

Jeong, D-I. and Kim, Y-O. 2005. ‘Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction’, Hydrological Processes, Vol 19(19), pp 3819 – 3835.

Karunasinghe, D.S.K. and Liong, S-Y. 2006. ‘Chaotic time series prediction with a global model: Artificial neural network’, Journal of Hydrology, Vol 323(1-4), pp 92 – 105.

Karunanithi, N., Grenney, W., Whitley, D. and Bovee, K. 1994. 'Neural Networks for River Flow Prediction', Journal of Computing in Civil Engineering, Vol 8(2), pp 201 – 220.

Kessler, E., and Neas, B. 1994. 'On correlation, with applications to the radar and raingage measurement of rainfall', *Atmospheric Research*, Vol 34(1-4), pp 217 – 229.

Kitanidis, P.K. and Bras, R.L. 1980. 'Real-Time Forecasting With a Conceptual Hydrologic Model: 2. Application and Results', Water Resources Research, Vol 16(6), pp 1034 – 1044.

Khalil, M., Panu, U.S. and Lennox, W.C. 2001. 'Groups and neural networks based streamflow data infilling procedures', Journal of Hydrology, Vol 241(3-4), pp 153 – 176.

Klemes, V. 1986. 'Operational testing of hydrological simulation models', Hydrological Sciences Journal, Vol 31(3), pp 13 – 24.

Krause P., Boyle, D.P. and Bäse, F. 2005. 'Comparison of different efficiency criteria for hydrological model assessment' in Krause, P., Bongartz, K. and Flügel, W-A. (eds) Advances in Geosciences Vol 5: Proceedings of the 8th Workshop for Large Scale Hydrological Modelling - Oppurg 2004, pp 89 – 97.

Kumar, A.R.S. Sudheer, K.P. Jain, S.K. and Agarwal, P.K. 2005. 'Rainfall-runoff modelling using artificial neural networks: comparison of network types', Hydrological Processes, Vol 19, pp 1277 – 1291.

Lauzon, N., Anctil, F. and Baxter, C.W. 2006. 'Classification of heterogeneous precipitation fields for the assessment and possible improvement of lumped neural network models for streamflow forecasts', Hydrology and Earth System Sciences Discussions, Vol 3(1), pp 201 – 227.

Legates, D.R., and Davis, R.E. 1997. **'**The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches**',** *Geophysical Research Letters*, Vol 24(18), pp 2319 **–** 2322.

Legates, D.R. and McCabe, G.J. 1999. 'Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation', Water Resources Research, Vol 35(1), pp 233 – 241.

Lin, G-F. and Chen, L-H. 2004. 'A non-linear rainfall-runoff model using radial basis function network', Journal of Hydrology, Vol 289(1-4), pp 1 – 8.

Lin, G-F. and Chen, L-H. 2005. 'Time series forecasting by combining the radial basis function network and self-organizing map', Hydrological Processes, Vol 19(10), pp 1925 – 1937.

Liong, S-Y., Lim, W-H. and Paudyal, G. 2000. 'River Stage Forecasting in Bangladesh: Neural Network Approach', Journal of Computing in Civil Engineering, Vol 14(1), pp 1 – 8.

Lorrai, M. and Sechi G. 1995. 'Neural nets for modelling rainfall-runoff transformations', Water Resources Management, Vol 9(4), pp 299 – 313.

Maier, H.R. and Dandy, G.C. 1996. 'The use of artificial neural networks for the prediction of water quality parameters', Water Resources Research, Vol 32(4), pp 1013 – 1022.

Maier, H.R. and Dandy, G.C. 2000. 'Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications', Environmental Modelling & Software, Vol 15(1), pp 101 – 124.

Makridakis, S. 1993. 'Accuracy measures: theoretical and practical concerns', International Journal of Forecasting, Vol 9(4), pp 527 – 529.

Nash, J.E. and Sutcliffe, J.V. 1970. 'River flow forecasting through conceptual models 1: A discussion of principles', Journal of Hydrology, Vol 10(3), pp 282 – 290.

NERC. 1975. *Flood Studies Report Vol. 1 - Hydrological Studies*. Natural Environment Research Council, London, UK.

Paik, K., Kim, J.H., Kim, H.S. and Lee, D.R. 2005. 'A conceptual rainfall-runoff model considering seasonal variation', Hydrological Processes, Vol 19(19), pp 3837 – 3850.

Pearson, K. 1896. 'Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia', Philosophical Transactions of the Royal Society of London Series A, Vol 187, pp 253 – 318.

Pebesma, E.J., Switzer, P. and Loague, K. 2005. 'Error analysis for the evaluation of model performance: rainfall-runoff event time series data', Hydrological Processes, Vol 19(8), pp 1529 – 1548.

Rajurkar, M.P., Kothyari, U.C. and Chaube, U.C. 2004. 'Modeling of the daily rainfall-runoff relationship with artificial neural network', Journal of Hydrology, Vol 285(1-4), pp 96 – 113.

Riad, S., Mania, J., Bouchaou, L. and Najjar, Y. 2004. 'Predicting catchment flow in a semi-arid region via an artificial neural network technique', Hydrological Processes, Vol 18(13), pp 2387 – 2393.

Refsgaard, J.C. and Henriksen, H.J. 2004. 'Modelling guidelines – terminology and guiding principles', Advances in Water Resources, Vol 27(1), pp 71 – 82.

Refsgaard, J.C. Henriksen, H.J. Harrar, W.G. Scholten, H. and Kassahun, A. 2005. 'Quality assurance in model based water management – review of existing practice and

outline of new approaches', Environmental Modelling and Software, Vol 20, pp1201 – 1215.

Schwarz, G. 1978. 'Estimating the dimension of a model', Annals of Statistics, 6(2), 461 – 464.

See, L. and Abrahart, R.J. 1999. 'Multi-model data fusion for hydrological forecasting', Proceedings of the 4th International Conference on Geocomputation, Fredericksburg, Virginia, USA, 25-28 July.

Seibert, J. 2001. 'On the need for benchmarks in hydrological modelling', Hydrological Processes, Vol 15(6), pp 1063 – 1064.

Shamseldin, A.Y. 1997. 'Application of a neural network technique to rainfall-runoff modelling', Journal of Hydrology, Vol 199(3-4), pp 272 – 294.

Singh, V.P. and Woolhiser, D.A. 2002. 'Mathematical Modeling of Watershed Hydrology', Journal of Hydrologic Engineering, Vol 7(4), 270 – 292.

Stephenson, D. 1979. 'Direct optimization of Muskingum routing coefficients', Journal of Hydrology, Vol 41(1-2), pp 161 – 165.

Sudheer, K.P., Nayak, P.C. and Ramasastri, K.S. 2003. 'Improving peak flow estimates in artificial neural network river flow models', Hydrological Processes, Vol 17(3), pp 677 – 686.

Supharatid, S. 2003. 'Application of a neural network model in establishing a stage-discharge relationship for a tidal river', Hydrological Processes, Vol 17(15), pp 3085 – 3099.

Tang, Y. Reed, P. and Wagener, T. 2006. 'How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration? ', Hydrology and Earth System Sciences, Vol. 10(2), pp 289 – 307.

Teegavarapu, R. and Elshorbagy, A. 2005. 'Fuzzy set based error measure for hydrologic model evaluation', Journal of Hydroinformatics, Vol 7(3), pp 199 – 207.

Wagener, T. and McIntyre, N. 2005. 'Identification of rainfall-runoff models for operational applications', Hydrological Sciences Journal, Vol 50(5), pp 735 – 751.

Watts, G. 1997. 'Hydrological Modelling in Practice', in Contemporary Hydrology: Towards holistic environmental science, Wilby, R.L. (ed), John Wiley, UK. pp 151 – 193.

Wilmott, C.J. 1981. 'On the validation of models', Physical Geography, Vol 2(2), pp 184 – 194.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J. and Rowe, C.M. 1985. 'Statistics for the evaluation and comparison of models', *Journal of Geophysical Research*, Vol 90(C5), pp 8995 – 9005**.**

WMO. 1986. *Intercomparison of models of snowmelt runoff*. Operational Hydrology Report No. 23, WMO Publ. No. 646, World Meteorological Organization, Geneva, Switzerland.

Yapo, P.O., Gupta, H.V. and Sorooshian, S. 1996. 'Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data', Journal of Hydrology, Vol 181(1-4), pp 23 – 48.

Yu, P-S. and Yang, T-C. 2000. 'Fuzzy multi-objective function for rainfall-runoff model calibration', Journal of Hydrology, Vol 238(1-2), pp 1 – 14.

Eq.	Metric	Example Studies
1*	AME	None found
2*	PDIFF	None found
3	MAE	Chang *et al*. (2004); Chen *et al*. (2006); Dawson *et al*. (2006a); Karunasinghe and Liong (2006); Liong *et al.* (2000); Supharatid (2003)
4*	ME	Chang *et al*. (2001)
5*	RMSE	Alvisi *et al*. (2006); Bowden *et al*. (2002); Campolo *et al*. (2003); Corani and Guariso (2005); De Vos and Rientjes (2005); Lauzon *et al*. (2006); Sudheer *et al.* (2003); Wagener and McIntyre (2005)
6	R4MS4E	Baratti *et al*. (2003); Cannas et al. (accepted)
7	AIC	Chibanga *et al*. (2003); Hsu *et al*. (1995)
8	BIC	Chibanga *et al*. (2003); Hsu *et al*. (1995)
9*	NSC	None found
10	RAE	None found
11	PEP	Lin and Chen (2004); Measured on annual basis: Sudheer *et al*. (2003)
12	MARE	Riad *et al*. (2004); Jain and Srinivasulu (2004; 2005)
13	MdAPE	Lin and Chen (2005)
14	MRE	Dawson *et al*. (2006a); Measured on a seasonal basis: Jeong and Kim (2005)
15	MSRE	Dawson *et al*. (2002); Dawson *et al*. (2006a)
16	RVE	Expressed as a percentage: Lin and Chen (2004); Rajurkar *et al*. (2004); Yapo *et al*. (1996); Yu and Yang (2000)
17	RSqr	Giustolisi and Laucelli (2005); See and Abrahart (1999)
18	CE	Boughton (2006); Coulibaly *et al*. (2000); Dawson *et al*. (2006a); De Vos and Rientjes (2005); Kumar *et al*. (2005); Shamseldin (1997);
19	IoAd	None found
20	PI	Anctil *et al*. (2004); Chang *et al*. (2004); Chen *et al*. (2006); DeVos and Rientjes (2005); Gaume and Gosset (2003); Jain and Srinivasulu (2005); Lauzon *et al*. (2006);
* used by the National Weather Service to calibrate the SAC-SMA Model (Gupta *et al*. 1998). ME and RMSE are expressed in terms of *daily measures* by the National Weather Service but the reported equations can be applied to different temporal periods.		

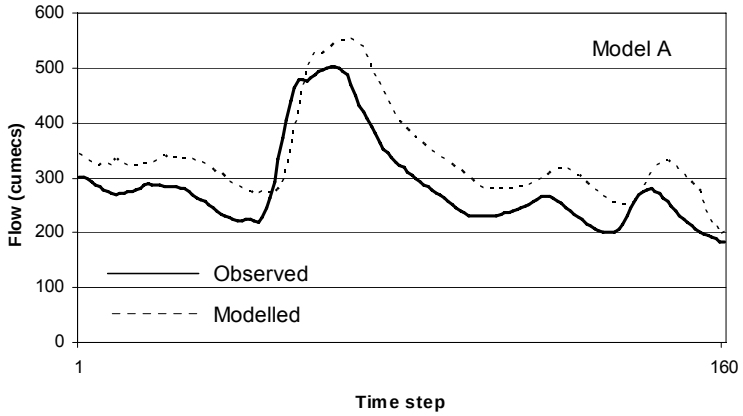
Table 1 Evaluation metrics used in recent reported hydrological modelling studies.

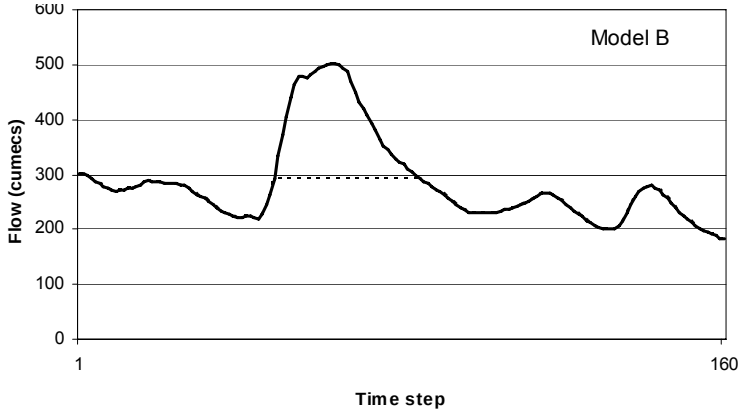
Statistic	Best	Worst	Good	Satisfactory	Poor
AME	0	∞	DD	DD	DD
PDIFF	0	±∞	DD	DD	DD
MAE	0	∞	DD	DD	DD
ME	0	±∞	DD	DD	DD
RMSE	0	∞	DD	DD	DD
R4MS4E	0	∞	DD	DD	DD
NSC	0	*n*	DD	DD	DD
RAE	0	∞	DD	DD	DD
PEP	0	±∞	DD	DD	DD
MARE	0	∞	DD	DD	DD
MdAPE	0	∞	DD	DD	DD
MRE	0	±∞	DD	DD	DD
MSRE	0	∞	DD	DD	DD
RVE	0	±∞	DD	DD	DD
RSqr	1	0	≥0.85	≥0.7	<0.7
IoAd	1	0	≥0.9	≥0.8	<0.8
CE	1	-∞	≥0.9	≥0.8	<0.8
PI	1	-∞	DD	>0	≤0

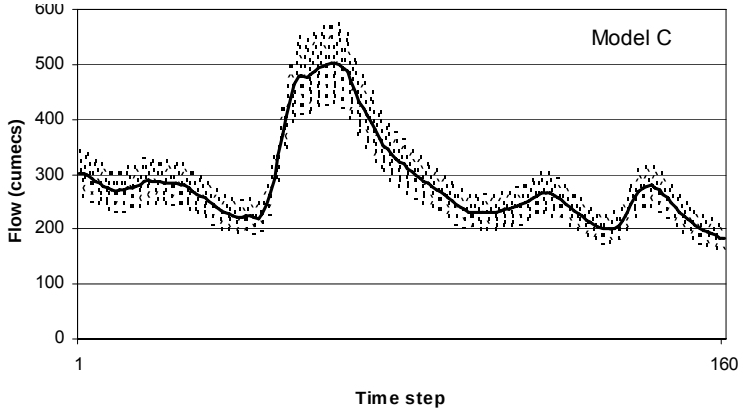
Table 2 Summary of statistical measures (DD is data dependent)

Statistic	Model A	Model B	Model C	Model D
AME	118	211	75.45	**75.25**
PDIFF	-50	202	-74.3	**0**
MAE	56.8875	**27.5688**	42.8728	45.4266
ME	-50.4500	27.5688	**-0.0534**	-45.4266
RMSE	61.2389	67.4775	**44.5832**	52.5474
R4MS4E	68.0833	111.6074	**48.6038**	57.5843
NSC	5	**1**	160	**1**
RAE	0.9677	**0.4690**	0.7293	0.7728
PEP	-9.9404	40.1590	14.7714	**0**
MARE	0.2090	**0.0612**	0.1500	0.1837
MdAPE	21.9879	**0.0000**	15.0000	25.0000
MRE	-0.1931	0.0612	**0.0000**	-0.1837
MSRE	0.0510	**0.0208**	0.0225	0.0447
RVE	-0.1765	0.0965	**-0.0002**	-0.1589
RSqr	0.8273	0.5604	0.7704	**0.9240**
IoAd	0.8714	0.6283	**0.9311**	0.8852
CE	0.4359	0.3151	**0.7010**	0.5847
PI	-43.6252	-53.1806	**-22.652**	-31.857

Table 3 Error measures computed on four hypothetical models


Model A
Observed
Modelled
Flow (cumecs)
Time step
0
100
200
300
400
500
600
1
160


Model B
Flow (cumecs)
Time step
0
100
200
300
400
500
600
1
160


Model C
Flow (cumecs)
Time step
0
100
200
300
400
500
600
1
160

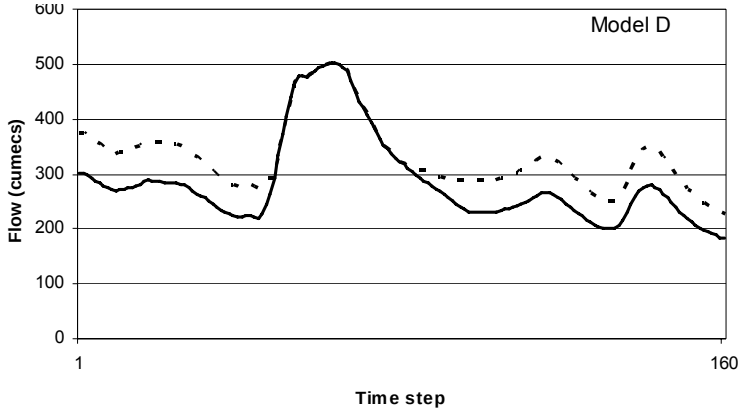
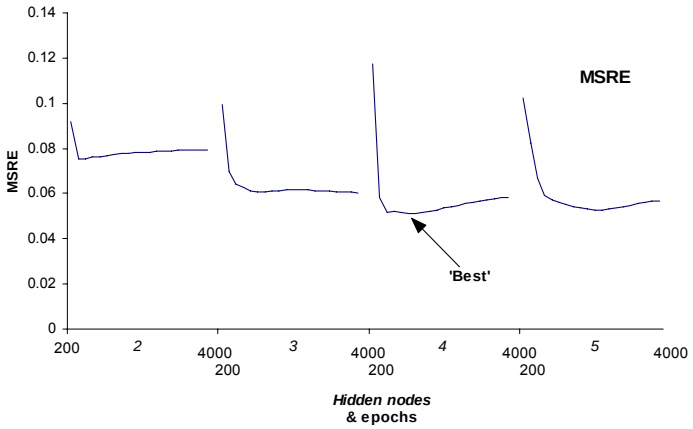
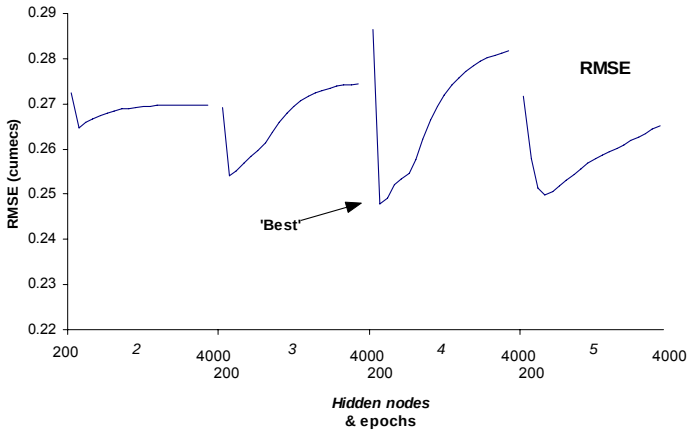
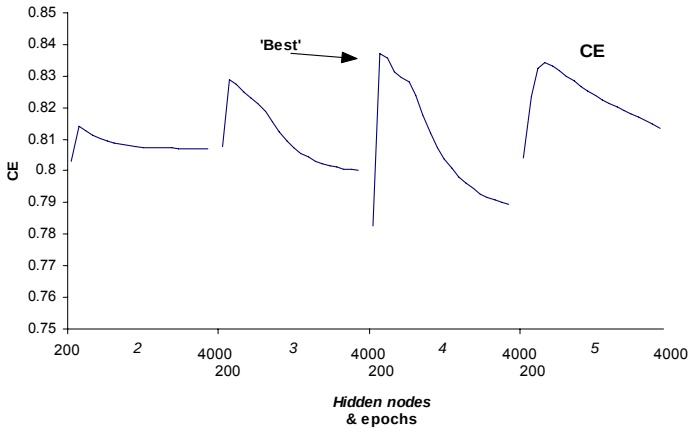

Model D
Flow (cumecs)
Time step
0
100
200
300
400
500
600
1
160

Figure 1 Four hypothetical river flow model outputs


MSRE
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
'Best'
200
2
4000
200
3
4000
200
4
4000
200
5
4000
Hidden nodes
& epochs


RMSE
RMSE (cumecs)
0.29
0.28
0.27
0.26
0.25
0.24
0.23
0.22
'Best'
200
2
4000
200
3
4000
200
4
4000
200
5
4000
Hidden nodes
& epochs


CE
0.85
0.84
0.83
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
'Best'
200
2
4000
200
3
4000
200
4
4000
200
5
4000
Hidden nodes
& epochs

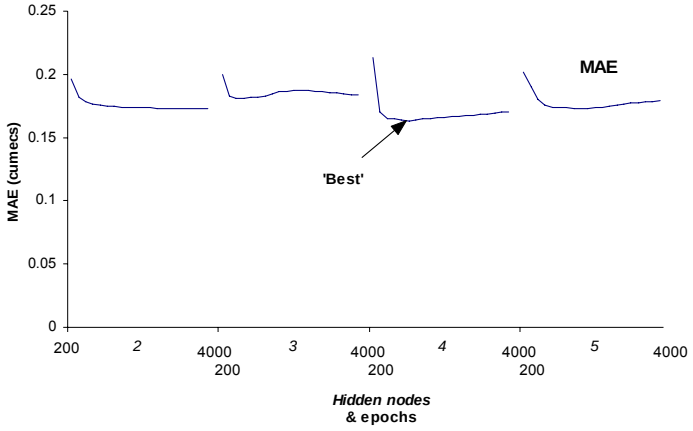

MAE
MAE (cumecs)
0.25
0.2
0.15
0.1
0.05
0
'Best'
200
2
4000
200
3
4000
200
4
4000
200
5
4000
Hidden nodes
& epochs

Figure 2 Error measures for different ANN models

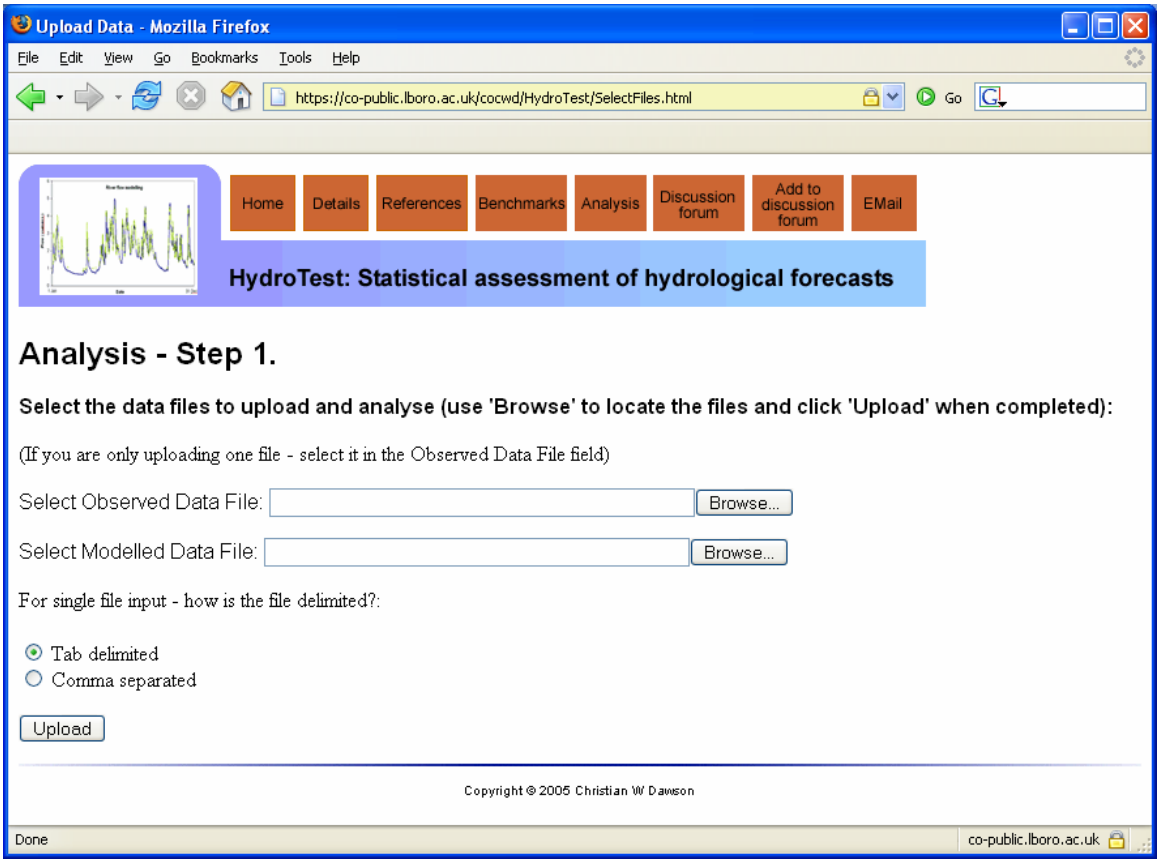


Figure 3 The *Analysis Step 1* screen

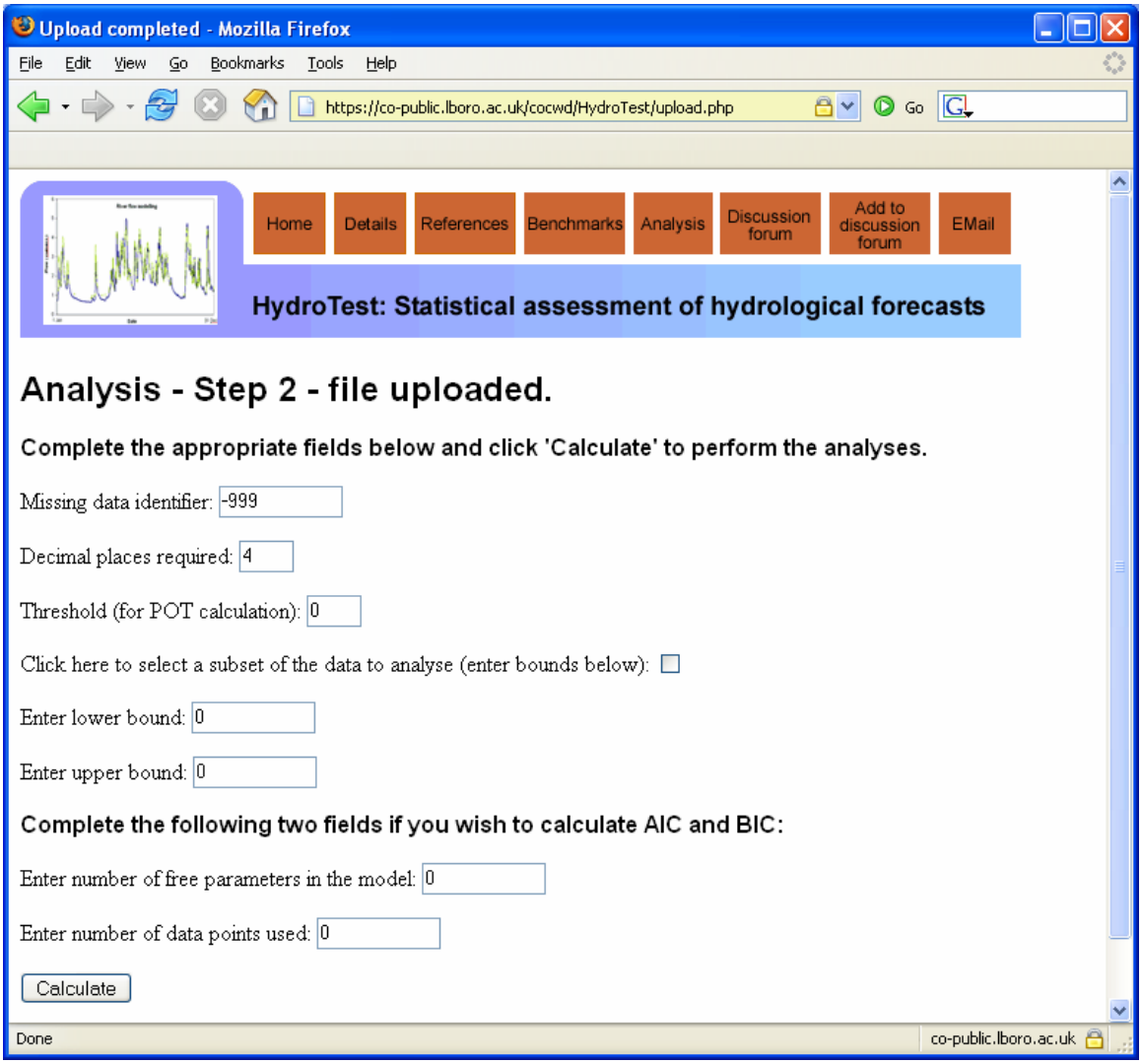


Figure 4 The *Analysis Step 2* screen

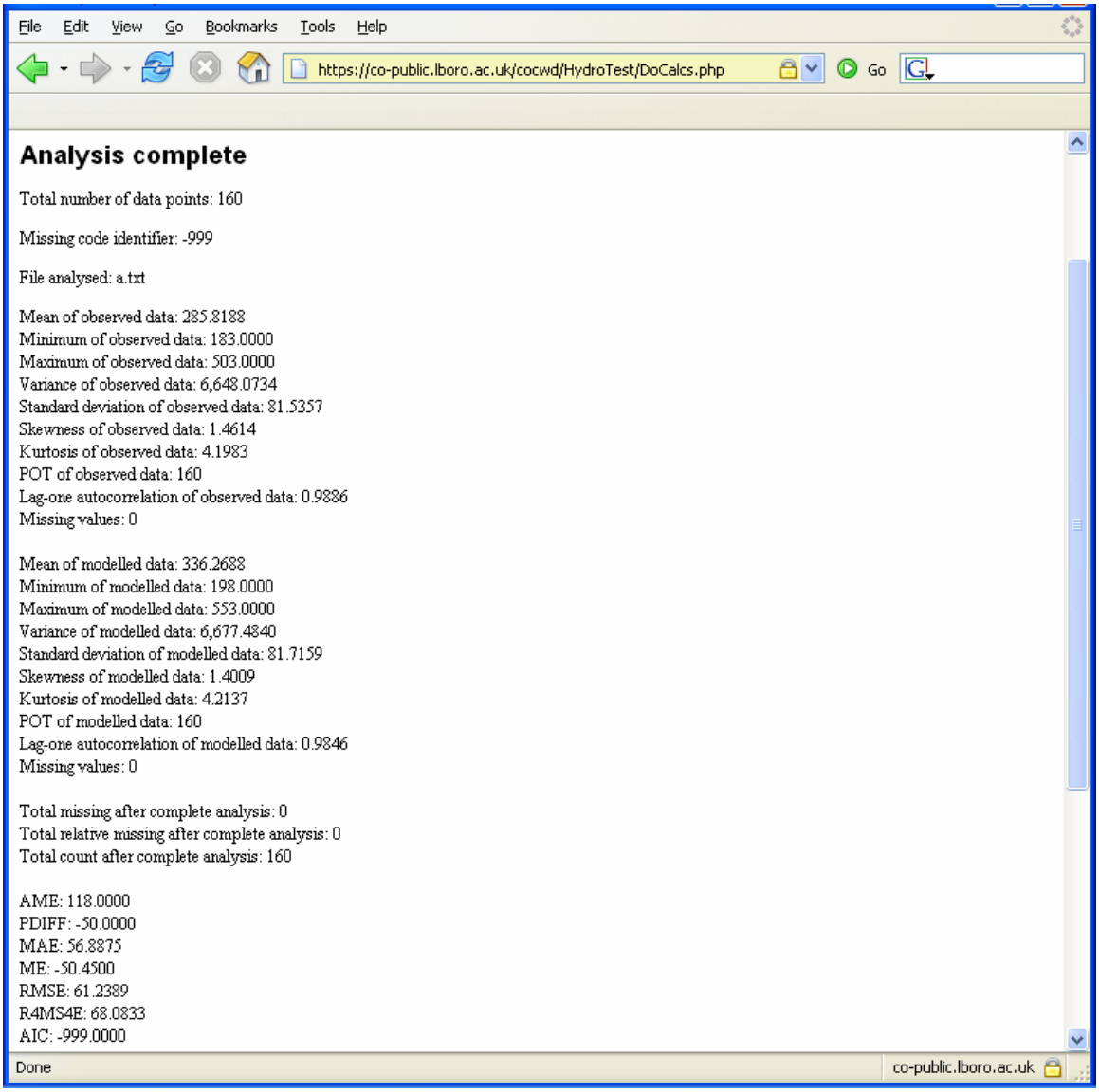


Figure 5 The results screen