The search for orthogonal hydrological modelling metrics: a case study of 20 monitoring stations in Colombia

E. Domínguez, C. W. Dawson, A. Ramírez and R. J. Abrahart

ABSTRACT

This paper presents a Colombian-based study on hydrological modelling metrics, arguing that redundancies and overlap in statistical assessment can be resolved using principal component analysis. Numerous statistical scores for optimal operator water level models developed at 20 hydrological monitoring stations, producing daily, weekly and ten-day forecasts, are first reduced to a set of five composite orthogonal metrics that are not interdependent. Each orthogonal component is next replaced by a single surrogate measure, selected from a set of several original metrics that are strongly related to it, and that in overall terms delivered limited losses with regard to 'explained variance'. The surrogates are thereafter amalgamated to construct a single measure of assessment based on Ideal Point Error. Depending on the forecast period, the use of three or four traditional metrics to deliver a combined evaluation vector, is the minimum recommended set of scores that is needed for analysis to establish the operational performance at a particular station in the gauging network under test.

Key words | hydrological modelling, ideal point error, orthogonal metrics, principal component analysis

E. Domínguez (corresponding author)
A. Ramírez
Departamento de Ecología y Territorio,
Pontificia Universidad Javeriana,
Transv. 4 # 42-00 Piso 8,
Bogotá, Colombia and CeiBA – Complejidad,
CRA. 1 Este # 18A - 70,
Bogotá, Colombia
E-mail: *e.dominguez@javeriana.edu.co*;
mathmodelling@gmail.com

C. W. Dawson
Department of Computer Science,
Loughborough University,
Loughborough LE11 3TU,
UK

R. J. Abrahart
School of Geography,
University of Nottingham,
Nottingham NG7 2RD,
UK

INTRODUCTION

Numerous statistical metrics are used by hydrologists to optimise or evaluate the output of their models: sometimes a single measure of assessment is used in isolation – for instance, during mean squared error model calibration procedures; and sometimes a trade-off involving two or more metrics is adopted. For example, two popular measures of model performance can be used that characterise different aspects of model behavior, the Mean Squared Error (*MSE*) criterion, which emphasises the fit to high flows (hydrograph peaks) and the Mean Squared Logarithmic Error (*MSLE*), which emphasises the fit to low flows (hydrograph recessions) (Pokhrel *et al.* in press). *MSE* measures the average size of the squared model residuals, whilst *MSLE* does the same, albeit after a log transformation has been applied. Most metrics will perform an assessment of the residual error occurring between observed and estimated time series records, providing quantitative descriptors in terms of absolute errors, relative errors or dimensionless coefficients. The decision on selecting an appropriate metric or set of metrics for a particular exercise is nevertheless complicated by the fact that a number of established measures exist. The modeller must be able to recognise precisely what each metric is sensitive to and understand how the ensuing output statistic is to be interpreted. The overall situation is further complicated by measurement similarities, sometimes delivering a restricted assessment of models due to inadequate selection of metrics and/or the incorporation of redundant descriptors. Moreover, the adoption of metrics that are interdependent can lead to poor analysis of model performance, and might result in wrong inferences or contradictory conclusions (Weglarczyk 1998; Shaefli & Gupta 2007).

Twenty-five conventional hydrological modelling metrics are currently available in HydroTest (http://www.hydrotest.org.uk): a standardised, open access website that performs the

doi: 10.2166/hydro.2010.116

required numerical calculations (Dawson *et al.* 2007, 2010). HydroTest does not attempt to provide a comprehensive set of evaluation metrics for all occasions: the current software set-up is instead oriented towards providing a practical tool for the assessment of quantitative modelling predictions in time and does not, at this point, assess vector processes or time- and space-distributed forecasts. To obtain a more comprehensive listing of forecasting performance metrics, one potential source is the World Weather Research Programme/World Climate Research Programme Joint Working Group on Forecast Verification Research (http://www.cawcr.gov.au/projects/verification/). The latest cohort of popular numerical hydrological descriptors in HydroTest are not truly orthogonal and in most cases possess insufficient discriminatory power to deliver a clear report on specific issues of hydrological interest. There is much overlap and limited exclusiveness with regard to the important hydrological qualities that are described, on a collective basis, under the general remit of an individual 'performance measure'. It is true that particular metrics will tend to focus on peak-flow, high-flow or low-flow situations, e.g. statistics that use squared error units will be biased towards an assessment of high(er) magnitude error; whilst statistics that compute relative measures will be biased towards an assessment of low(er) magnitude error. The main point, however, is that in most cases statistical assessment is being applied, on a global basis, to a continuum of responses computed across a complex series of events; something that will at best deliver a set of blurred results – a muddying of the water! Local snapshots such as measuring forecasting performance in terms of success on the 'highest flood event' are, in contrast, rather simple and not that representative of the overall situation. The other option would be to implement fitness testing on different partitions of a particular dataset: but such activities might lead to questions being asked about the rights and wrongs of selecting each particular subgroup. For a recent example of separating river flow series into almost independent quick- and slow-flow hydrograph periods to support the multi-criteria performance evaluation of rainfall–runoff models, see Willems (2009).

In Colombia, hydrological variability is the major source of uncertainty in water management processes, and such matters have a direct effect on the operational efficiencies of disaster prevention organisations and management services (Costa *et al.* 2005). To help address such issues, and provide better support for the government and private sectors, the Colombian Institute of Hydrology, Meteorology and Environmental Research (IDEAM: http://www.ideam.gov.co) has embarked upon a modernisation scheme for its hydrological forecasting services and procedures. IDEAM has focussed on three key points: (i) acquisition of modern hydrological instruments and recorders, (ii) enhancement of real-time data transmission technologies and (iii) the development and tuning of a practical national forecasting methodology.

In 2005, a sequence of projects was started to identify a feasible option for the core component of a real-time forecast system: the IDEAM 'Colombian Hydrometeorological Warning System' or CHWS (Mussy 2005; Rivera *et al.* 2005). In parallel, another project analysed infrastructure issues for the information fluxes that support the daily forecasts at IDEAM (Rojas 2006). Earlier modelling at IDEAM (Domínguez 2004; Rivera *et al.* 2004) has been very valuable, showing that a diversity of tools can be used within the proposed system, but most such projects lacked common objective metrics for model assessment, making it difficult to establish which of the proposed solutions was best suited as the core component of CHWS. In order to determine an appropriate assessment procedure for different solutions and situations the capability for water level forecasting was studied at 20 hydrological stations belonging to CHWS (Figure 1). Daily, weekly and ten-day forecasting models were developed for each station. To help overcome issues of assessment metric redundancy and interdependence, and to test for the notion of a set of orthogonal performance criteria, principal component analysis (PCA) is first applied to 60 sets of 22 different statistical metrics. Each set of metrics comprised performance indicators computed on outputs for one of the forecasting models. The top five principal components were thereafter scrutinised and a single surrogate metric, selected from a range of original candidate metrics, was chosen to represent each orthogonal component. The preferred surrogates are then combined, adopting a stepwise procedure, to deliver a single 'distance from best point' metric such that each station can be ranked one against another in relative terms. Further analysis is used to ascertain the number of surrogate inputs needed to provide a sufficient description of daily, weekly and ten-day forecasting models.

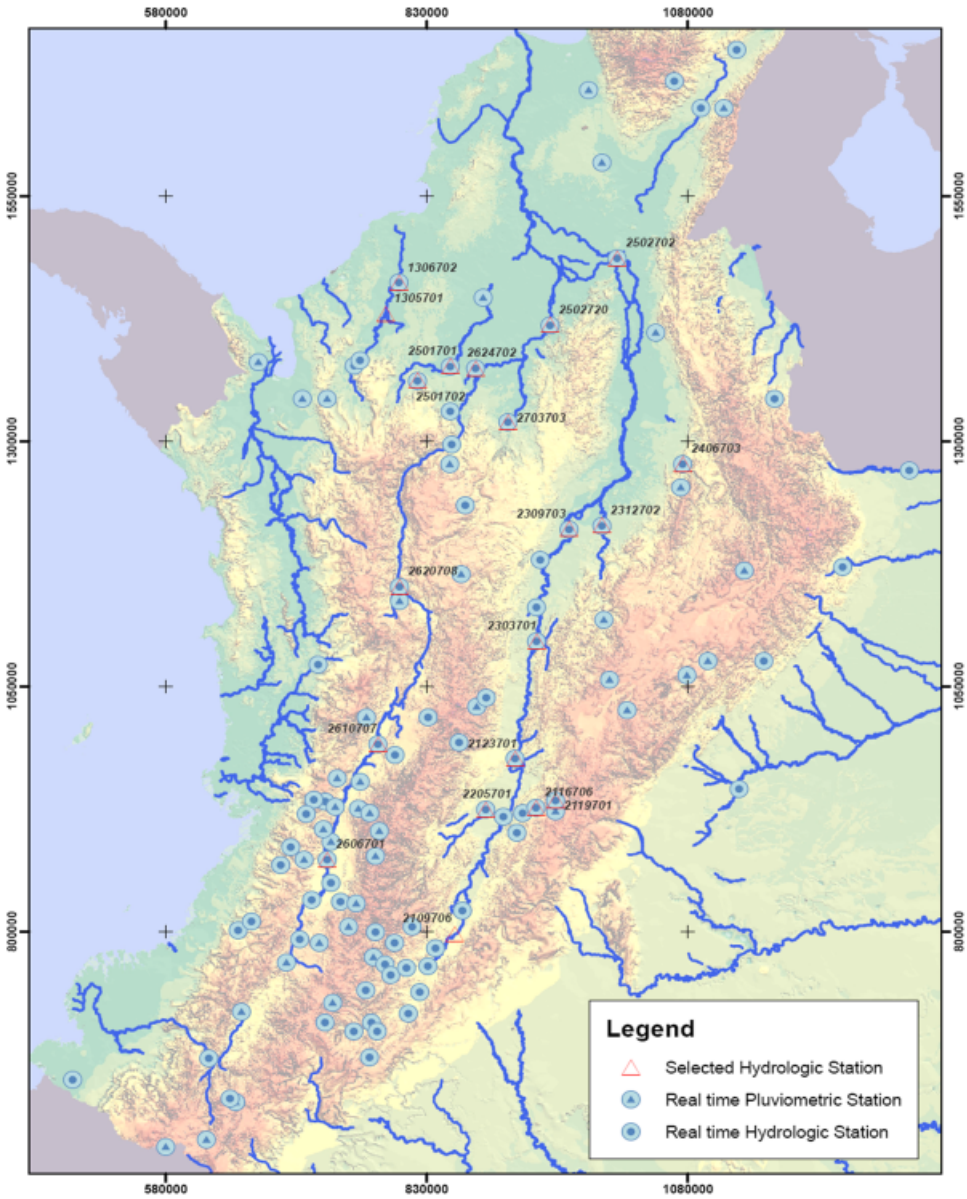

Figure 1 | The Colombian real-time hydrological network and selected forecast points.

EXPERIMENTAL SET-UP

The main objective of this work is to establish the relative operational forecasting capabilities at 20 hydrological stations that are collecting and transmitting water level data in real time. Forecasting assessment was performed on models of daily, weekly and ten-day accumulated water levels in five steps: (i) selection of mathematical forecasting operators for 60 models, (ii) assessment of modelling outputs using 22 traditional hydrological performance metrics provided in HydroTest, (iii) production of an orthogonal set of alternative performance metrics, (iv) selection of preferred surrogate metrics and (v) amalgamation of surrogate metrics into a single measure of overall success.

STUDY AREA AND DATASET

Hydrometeorological datasets were provided by IDEAM. Their hydrological real-time monitoring network has 45

active stations from which 20 forecast points were selected for evaluation. Stations were chosen taking into account different topographic, physiographic and hydrological conditions. The selected stations are listed in Table 1 and their locations are shown in Figure 1. The catchment areas for the selected hydrological stations vary from 132 to 161292 km^2, the altitudinal ranges vary from 14 to 5375 m above sea level and streamflow ranges from 4.5 to 4200 m^3/s (Table 1). From the selected hydrological stations, five years of daily water level data were obtained. The selected sites included forecasting points with short and long 'process memories' and had no gaps in their records. Weekly and ten-day time series comprised daily data totalled by seven- and ten-day periods. The autocorrelation radius, the time series lag with non significant autocorrelation coefficient, for daily water levels varied from one day of process memory for mountainous rivers to more than 60 days of memory for flat land rivers. The coefficient of variation (C_v), that represents the variability of daily water levels, shows some stations with very regular water level behaviour ($C_v \leq 0.2$) and others with highly irregular water level oscillations ($C_v > 0.2$). The lag-one autocorrelation coefficients range from 0.57 (Station 2116706: San Pablo–Cuinde River) to 0.96 (Station 1306702: Monteria–Sinu River). In general, the bigger the catchment area, the longer the 'process memory' and vice versa. The range of catchment areas studied is sufficient to ensure that a wide spectrum of hydrological variability is included. This catchment descriptor, however, is not able to fully describe the water level inertia at the different stations. Daily rainfall records were also available from local gauges for use as exogenous

Table 1 | Selected stations, catchments and water level characteristics

	Catchment characteristics						Water level characteristics (using hydrological local reference system)				
Hydrological station code	Area [kms^2]	Streamflow length, [kms]	Min elevation above sea level, [m]	Max elevation above sea level, [m]	Catchment average elevation about sea level, [m]	Discharge flow, [m^3/s]	Mean	Standard Deviation	Coefficient of Variation	Lag one autocorrelation coefficient	Autocorrelation radius, days
1305701	7974	208	24	3712	450	330	2.99	1.13	0.38	0.95	40
1306702	9246	255	14	3712	396	370	2.67	1.15	0.43	0.96	38
2104701	5586	148	695	4618	1898	220	1.75	0.40	0.23	0.76	2
2109707	15037	267	421	5375	1895	480	2.69	0.50	0.19	0.74	5
2116706	205	15	640	NA	1750	6.2	2.12	0.23	0.11	0.57	2
2119701	924	20	1780	4149	3314	19.5	2.15	0.43	0.20	0.62	1
2123701	44714	503	257	5375	1808	1120	2.57	0.96	0.38	0.84	5
2205701	7785	173	317	4581	2103	310	1.68	0.44	0.26	0.62	2
2303701	56054	660	170	5375	1738	1640	1.55	0.28	0.18	0.79	5
2309703	74190	820	104	5375	1558	2420	3.48	0.48	0.14	0.91	6
2312702	5361	183	99	3780	1210	265	2.60	1.02	0.39	0.87	37
2315703	87692	958	70	5375	1431	NA	2.30	1.07	0.46	0.96	60
2406703	21216	343	194	5340	2392	490	2.22	0.55	0.25	0.90	17
2501701	3960	172	36	3371	442	200	2.01	1.03	0.51	0.93	47
2501702	442	39	59	1841	300	25	2.43	0.32	0.13	0.72	4
2502702	161292	1107	24	NA	1290	4200	6.31	1.67	0.26	0.95	46
2601707	132	24	2895	NA	3500	4.5	2.94	1.09	0.37	0.93	20
2606701	8610	221	948	4636	1985	275	2.71	1.42	0.52	0.91	46
2620708	30973	600	555	5280	1850	860	2.51	0.76	0.30	0.94	20
2623704	37417	818	132	5280	1817	1200	2.33	0.56	0.24	0.92	21

predictors. The density of precipitation stations transmitting data in real time is twice as large as the density of hydrological stations; the former measure a wide range of meteorological variables, but only daily rainfall was retrieved.

FORECASTING TECHNIQUE

To obtain the initial forecasts, a forecast operator was selected, and relevant modelling outputs calculated for each gauging station and forecast period. The preferred method was based on previous findings which concluded that, in the Colombian real-time hydrological network and its hydrometeorological data fluxes, an optimal operator technique is the most suitable method to forecast daily, weekly and monthly water levels (Domínguez 2007b). This technique is computationally very efficient and simple to implement. Another attractive point is that the optimal operator technique is sufficiently flexible to assimilate a new predictor series when the stations delivering the calibrated inputs have gone offline (Domínguez *et al.* 2009, 2010).

The implementation of optimal operators requires four parameters: (i) amount of lag for exogenous and (ii) endogenous variables (n_{X_k} n_y), (iii) forecast horizon (lead time) T and (iv) length of parameterisation time window θ. These parameters are found by means of optimisation (exhaustive search). Usually, $\theta << N$ (where N is the time series length for $Y(t)$ and $X(t)$). The various pros and cons of selecting a particular forecasting procedure is not however the core argument of this paper; our interest is instead focused on methods of forecast capability assessment. The reported investigation, as a result, is limited to a consideration of one-time-step lead time solutions ($T=1$), meaning that each model was required to provide a forecast of the accumulated water level for the following day, week or ten-day period. For the purpose of this paper, the search domain for the optimal operator is reduced to the field of first-order (linear) polynomials and least-squared error optimisation procedures. For modelling greater lead times see Domínguez *et al.* (2009). To establish the required coefficients, we use only selected exogenous and endogenous data constrained inside a parameterisation window of length θ. For a better understanding of how this technique works, we present a graphical example in Figure 2. Here we show a linear optimal operator using one exogenous variable (rainfall). To produce a water level forecast for the moment $(t+T)$ with a lead time of $T=1$, we set a parameterisation window using data registered for the time interval [t-θ, t] and apply a linear regression algorithm to find the required coefficients. In this case, we only have one exogenous variable (so the k index can be omitted). The optimal operator is then a function of

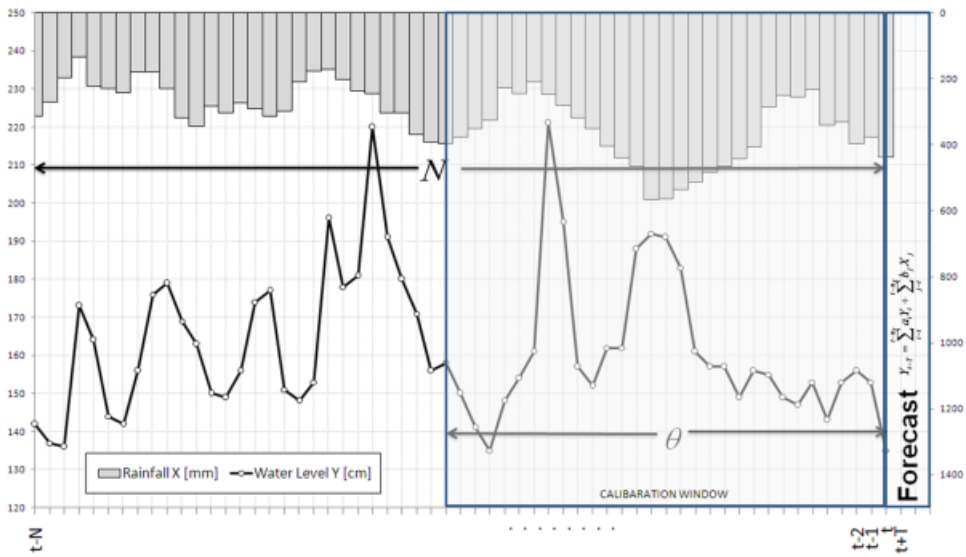

Figure 2 | Explaining the optimal operator forecast: N is the length of observed data, θ the calibration window length, X the exogenous predictor, Y the variable to be forecast and T the forecast lead time.

the type:

$$Y_{t+T} = a_t Y_t + a_{t-1} Y_{t-1} + a_{t-2} Y_{t-2} + \cdots + a_{t-n_y} Y_{t-n_y} + b_t X_t + b_{t-1} X_{t-1} + b_{t-2} X_{t-2} + \cdots + b_{t-n_X} X_{t-n_X} = \sum_{i=t}^{t-n_Y} a_i Y_i + \sum_{j=t}^{t-n_X} b_j X_j \quad (1)$$

Thus Y_{t+T} is expressed as a linear combination of lagged values for Y and X. Here, $n_Y + n_X$ represents the number of selected predictors that, together with the length θ, are optimised to find the optimal linear combination (OLC). To find OLC prior to the current forecast, we try $N-\theta$ forecasts, using historical data and varying θ and the number of predictors in an attempt to fulfil the constraint $S/\sigma_\Delta \leq 0.8$ (where S is root mean squared error, and σ_Δ is standard deviation, of forecast variable increments in the lead time interval). It should, however, be noted that only 25 of our 60 one-step-ahead models achieved the required standard. The trial forecast is performed using a moving window procedure (Figure 3) that runs from the trial forecast of the registered value $Y_{t-(N+\theta+1)}$ to the trial forecast of Y_t. Then, having established an optimal θ and the optimal lags of selected predictors (exogenous or endogenous), the water level Y_{t+T}

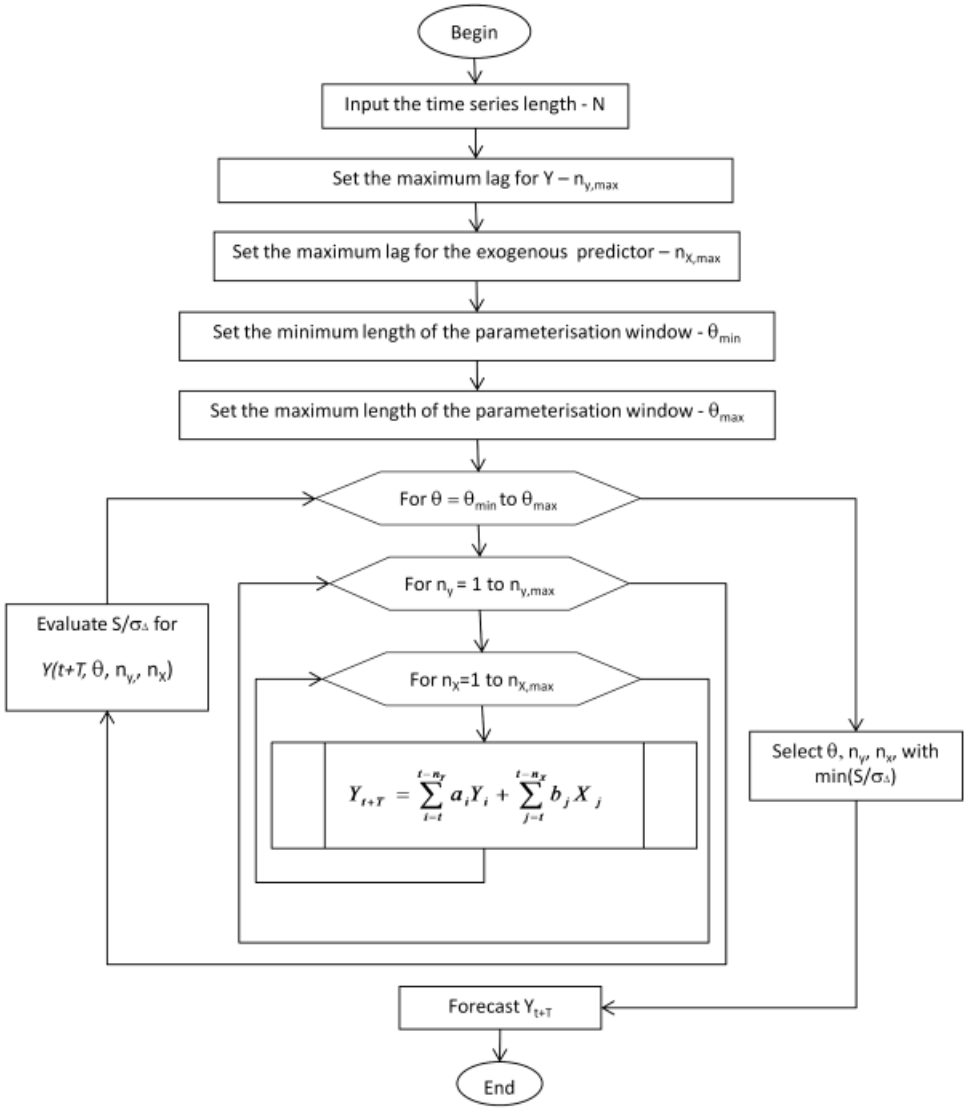

Figure 3 | Flowchart of the trial forecast moving window procedure for the case of one exogenous predictor.

can be forecasted. The procedure is revisited each time a forecast is issued.

This method is intended to be the core of a real-time forecasting system; its validation scheme thus differs from that which is normally applied due to the use of updating procedures. For short-term forecasts it is not necessary to guarantee a long-term validation and in such cases it is more important to have a good expectation of the forecast error at the next time step. Such an expectation is obtained by validating forecasts, during the interval arising between the last forecast and the next forecast, together with optimisation of the θ parameter and identification of optimal lags for selected predictors. This validation is possible because of the support provided by real-time data transmission about the forecasted hydrological variable and its exogenous predictors.

Despite the rapid growth of modelling techniques and data sources, problems related to a lack of standard terminology and modelling protocols still exists in hydrological forecasting. One issue is related to validation and performance assessment procedures. There is, for example, a conceptual discussion about when and where a model can be verified or validated (Oreskes *et al.* 1994; Beven 2002; Refsgaard & Henriksen 2004). In addition, there is no agreement about which performance criteria should be used to decide if a model structure is suitable for inclusion as a tool for hydrological forecasting. Some efforts towards the establishment of a standard modelling protocol include Domínguez (1998, 2007*a*), Van Waveren *et al.* (1999) and Refsgaard & Henriksen (2004). Model developers, moreover, do not provide consistent or standard quantitative evaluation statistics and it is sometimes difficult for readers or users to determine how well a particular model reproduces the observed dataset or how well a model compares with other models (ASCE 1993; Legates & McCabe 1999).

HydroTest was used in the reported investigation to help overcome such matters: providing a set of transparent, objective and consistent assessment metrics. This tool delivers 25 measures of performance which are evaluated independently from the modeller and ensures the replication of performance assessment in the case of external audits. It also calculates the mean, variance, standard deviation, skewness, kurtosis and lag-one autocorrelation coefficient of observed and modelled datasets. HydroTest is still growing in popularity; towards the end of 2010 it had over 750 registered users from over 25 countries worldwide. Several studies in which HydroTest has been used for analysis have already been published. Examples include Dawson *et al.* (2006), Abrahart *et al.* (2007) and Napolitano *et al.* (2010). The performance metrics used in HydroTest are classified into three groups: absolute, relative and dimensionless. Denoting observed and forecasted values as Q_i and $\hat{Q}_i$ and the means for observed and forecast values as $\overline{Q}$ and $\hat{Q}$, respectively, the HydroTest metrics that were used in this study are as follows:

(i) Absolute metrics

Absolute maximum error

$$\mathrm{AME} = \max_{i=1}^{n} \left| Q_i - \hat{Q}_i \right| \tag{2}$$

Peak difference

$$\mathrm{PDIFF} = \max_{i=1}^{n}(\mathrm{Q_i}) - \max_{i=1}^{n}(\hat{\mathrm{Q}}_i) \tag{3}$$

Mean absolute error

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} \left| \mathrm{Q_i} - \hat{\mathrm{Q}}_i \right| \tag{4}$$

Mean error

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

Root mean squared error

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{Q_i} - \hat{\mathrm{Q}}_i)^2}{n}} \tag{6}$$

Fourth root mean quadrupled error

$$\mathrm{R4MS4E} = \sqrt[4]{\frac{\sum_{i=1}^{n} (\mathrm{Q_i} - \hat{\mathrm{Q}}_i)^4}{n}} \tag{7}$$

Number of sign changes

$$\mathrm{NSC} = \text{Number of sign changes for residuals} \tag{8}$$

Mean squared logarithmic error

$$\mathrm{MSLE} = \frac{1}{n}\sum_{i=1}^{n}(\ln Q_i - \ln \hat{Q}_i)^2 \tag{9}$$

Mean squared derivative error

$$\mathrm{MSDE} = \frac{1}{n-1}\sum_{i=1}^{n}((Q_i - Q_{i-1}) - (\hat{Q}_i - \hat{Q}_{i-1}))^2 \tag{10}$$

(ii) Relative metrics

Relative absolute error

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n}\left|Q_i - \hat{Q}_i\right|}{\sum_{i=1}^{n}\left|Q_i - \overline{Q}_i\right|} \tag{11}$$

Percent error in peak

$$\mathrm{PEP} = 100 \times \frac{\max_{i=1}^{n}(Q_i) - \max_{i=1}^{n}(\hat{Q}_i)}{\max_{i=1}^{n}(Q_i)} \tag{12}$$

Mean absolute relative error

$$\mathrm{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Q_i - \hat{Q}_i\right|}{Q_i} \tag{13}$$

Median absolute percentage error

$$\mathrm{MdAPE} = \mathrm{median}_{i=1}^{n}\left(\frac{\left|Q_i - \hat{Q}_i\right|}{Q_i}\right)x100 \tag{14}$$

Mean relative error

$$\mathrm{MRE} = \frac{1}{n}\sum_{i=1}^{n}\frac{(Q_i - \hat{Q}_i)}{Q_i} \tag{15}$$

Mean squared relative error

$$\mathrm{MSRE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_i - \hat{Q}_i}{Q_i}\right)^2 \tag{16}$$

Relative volume error

$$\mathrm{RVE} = \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)}{\sum_{i=1}^{n}Q_i} \tag{17}$$

(iii) Dimensionless metrics

Coefficient of determination

$$\mathrm{RSqr} = \left(\frac{\sum_{i=1}^{n}(\hat{Q}_i - \tilde{Q})(Q_i - \overline{Q}_i)}{\sqrt{\sum_{i=1}^{n}(Q_i - \overline{Q})^2\sum_{i=1}^{n}(\hat{Q}_i - \tilde{Q})^2}}\right)^2 \tag{18}$$

Index of agreement

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

Coefficient of efficiency

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

Persistence index

$$\mathrm{PI} = 1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - Q_{i-1})^2} \tag{21}$$

Inertia root mean squared error

$$\Delta_i = Q_i - Q_{i-1}$$
$$\overline{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\Delta_i \tag{22}$$

$$\sigma_\Delta = \sqrt{\frac{\sum_{i=1}^{n}(\Delta_i - \overline{\Delta})^2}{n-1}}$$

$$\text{IRMSE} = \frac{RMSE}{\sigma_\Delta}$$

$$\alpha = \sigma_{\hat{Q}}/\sigma_Q$$

$$\beta = \tilde{Q}/\overline{Q}$$

Kling-Gupta efficiency

$$\begin{aligned} &r = \text{linear correlation coefficient} \\ &\text{ED} = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \\ &\text{KGE} = 1 - \text{ED} \end{aligned} \tag{23}$$

RESULTS AND DISCUSSION

Principal Component Analysis

PCA is a statistical method that is used to transform a set of original variables into a smaller set of uncorrelated, derived variables, called components – with the latter being designed to account for most of the variability (information) that is contained in the original material (Dunteman 1989). PCA, in the context of this paper, is applied to the 22 hydrological modelling statistics that were computed for each of the 60 forecast cases, in order to identify a smaller set of orthogonal measures. From this analysis, five principal components were identified having eigenvalues greater than unity – a general rule-of-thumb for selecting which principal components to retain. The first five principal components in this instance accounted for 91.1% of the variance; put another way, the five new variables could account for most of the information that was originally covered by 22 metrics. Table 2 presents a statistical summary of the first five principal components: magnitude, explained variance and cumulative explained variance are reported.

Following the identification of five principal components, the statistical procedure continues by rotating the principal components using the varimax criterion (one of several approaches that can be used) such that the sum of the variances of the squared loadings in each column of the loading matrix is maximised (Kaiser 1958). In other words, a new set of orthogonal coordinate axes are produced with large or small loadings for each of the original variables on it (Dunteman 1989). This procedure helps to identify which of the original variables are contributing most to each of the (now rotated) principal components. Table 3 presents the varimax rotated component loadings (correlations) of the first five principal components for each of the original hydrological modelling statistics.

Table 2 | First five principal components (unrotated)

Principal component	Magnitude	Explained variance	Cumulative explained variance
1	10.56	0.480	0.480
2	4.66	0.212	0.692
3	2.34	0.106	0.798
4	1.34	0.061	0.859
5	1.15	0.052	0.911

Those variables exhibiting strong correlations with each of the principal components are highlighted in Table 3. In some cases (the first principal component, for example) a number of variables are similarly correlated with the component. By identifying those variables strongly associated with each component it is possible to derive some kind of meaning (representation) for each of the five principal components.

The first principal component is strongly correlated with AME, MAE, RMSE, R4MS4E and MSDE. Four of the five error measures provide some indication of the accuracy of the model at each data point, the exception being MSDE. The first principal component is perhaps giving some indication of the general accuracy of the models.

The second principal component is strongly correlated with RAE, RSqr, IoAd, CE and KGE. The five error measures are perhaps an indication of a model's representation of the shape of the observed hydrograph, implying that the second component measures how closely a model simulates the overall profile of rises and falls in a series.

The third component shows strong correlation with two statistics – ME and RVE. RVE is calculated in a similar manner to ME except that the resultant measure is divided by the total observed record. This component perhaps indicates the total volume error of the model.

Table 3 | Varimax rotated component loadings (correlations)

Component	1	2	3	4	5
AME	**0.944**	−0.212	−0.032	0.049	−0.029
PDIFF	0.652	−0.411	−0.090	−0.193	0.431
MAE	**0.963**	−0.181	−0.030	−0.008	−0.082
ME	−0.130	0.177	**0.926**	−0.62	−0.067
RMSE	**0.967**	−0.182	−0.030	0.000	−0.095
R4MS4E	**0.968**	−0.188	−0.019	0.020	−0.099
NSC	−0.430	0.071	−0.050	0.118	**0.795**
RAE	0.206	**−0.946**	−0.060	0.076	0.017
PEP	0.199	−0.681	0.013	−0.050	0.531
MARE	0.774	−0.128	−0.273	0.470	0.193
MdAPE	0.697	−0.200	−0.240	0.527	0.218
MRE	−0.489	0.123	0.763	−0.213	−0.219
MSRE	0.782	0.057	−0.237	0.357	0.157
RVE	−0.062	0.074	**0.952**	−0.105	0.062
RSqr	−0.119	**0.976**	0.033	0.004	−0.005
IoAd	−0.085	**0.969**	0.022	0.050	−0.020
CE	−0.119	**0.976**	0.036	0.000	0.017
PI	−0.220	0.647	0.070	**−0.613**	0.039
MSLE	0.772	0.025	−0.167	0.467	0.184
MSDE	**0.909**	−0.173	−0.148	−0.170	0.074
IRMSE	0.204	−0.641	−0.082	**0.620**	−0.026
KGE	−0.100	**0.963**	0.038	0.044	−0.077

Bold = strong correlation

The fourth component is strongly correlated with PI and IRMSE which are measures based on the performance of the model against a simple model using observed data from the previous time step. This component is perhaps a measure of how well the model performs compared with a naïve one-step-ahead model.

The final component is strongly correlated with NSC. This is a fairly unique measure that merely counts the number of sign changes of the residuals of a model. NSC captures information that none of the other statistics remotely addresses. It is not surprising, therefore, that PCA has identified this final component, since it captures information in the dataset that nothing else contains.

Having identified five key principal components from the dataset and identified which statistics are strongly correlated with each component it is possible to select a subset of variables to represent these key components.

As a starting point, Figure 4 shows the rotated loadings of each of the statistics on the first two principal components. It is clear from this diagram that a number of statistics exhibit similar qualities for 69% of the total variation of the original set. For example, MSLE, MSRE, MAE, MARE, MdAPE, AME, RMSE and R4MS4E have similar loadings for the first two principal components. RSqr, CE, IoAd and KGE also exhibit similar qualities, as do IRMSE, PEP and RAE. Each group is circled in Figure 4. Although ME and RVE appear to contribute little to the first and second components, they are both strongly correlated with the third component – which is orthogonal to components one and two in this diagram (they would appear close to one another in a 3D plot on the z axis). Therefore, ME and RVE have also been circled in Figure 4. One could use a principal component plot such as this to identify representative statistics for each component but other approaches

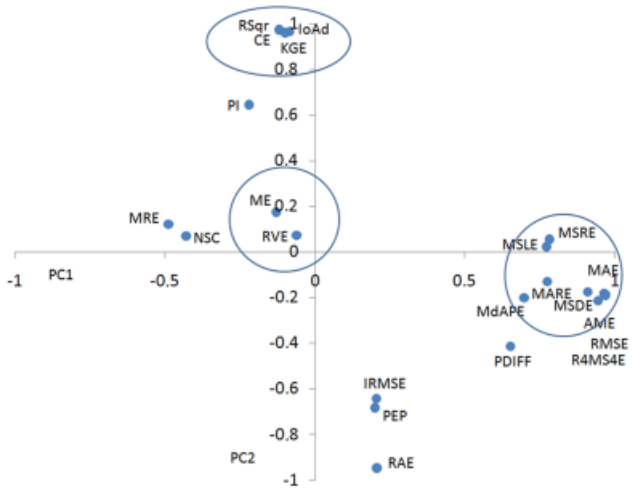

Figure 4 | Rotated loadings of all statistics on first two principal components.

exist – a discussion of such techniques can be found in Jolliffe (1986).

The initial approach adopted here is; for each principal component in turn (starting from the principal component with the highest variance and working downwards) select the variable with the highest loading on that component. Repeat this process for each of the components but note that a variable can only be selected once. Applying this procedure to our data results in the following subset of selected variables: R4MS4E, CE, RVE, IRMSE and NSC. One can calculate the amount of variation this subset explains by regressing each of the discarded statistics with the retained variables. In this case the subset of selected variables accounts for 86% of the variance of the original 22 metrics (not too different from the 91% explained by the first five principal components).

Unfortunately this subset contains some relatively underused or little-recognised statistics. With this problem in mind the loadings of the principal components were re-examined and an alternative set of five representative metrics was selected. In this case the statistics were selected based, not only on strong correlations with the corresponding principal components, but also on their popularity within the general body of literature. In this case RMSE, RSqr, ME, PI and PEP were selected. PEP is selected over NSC because NSC is a metric that does not provide any real indication on the accuracy of a model. A perfect model would result in an NSC value of zero, but likewise so might a wholly inaccurate model. In addition, PEP can be used in the calculation of an Ideal Point Error (see below), whereas NSC cannot, since it is not possible to define a perfect score. The variance this new subset explains of the original 22 metrics is 85% – virtually identical to the optimal subset identified earlier. In this case, however, the subset contains metrics with which most modellers should be familiar.

While the selection of these five representative statistics has been made on a combination of objective (based on component loadings) and subjective (based on an understanding of their popularity within the literature) criteria, the process of identifying these statistics highlights an important issue. The fact remains that a number of reported statistics presented in the literature overlap considerably. The process of selection applied here has shown that representative statistics should, in some way, be orthogonal to one another. An optimal set of RMSE, RSqr, ME, PI and PEP has been selected in this case, but we could just as easily have selected another five, providing each statistic is orthogonal and provides some representation of each of the generic error measures (principal components). Nevertheless, it should be remarked that, if we consider the information loading of each component, using the first three components will not be that different from using the first five components because of the low contributions of the fourth and fifth components. The results presented here show a clear clustering of the metrics and it would be up to the modeller to select an appropriate subset from these clusters.

Ideal point error

Taking the five selected statistics it is possible to combine them to produce a single error measure for comparing each of the models in the dataset. An *Ideal Point Error* (IPE) measurement is calculated by identifying the ideal point in five-dimensional space that each model should be evaluated against (based on Elshorbagy *et al.* (2010)). IPE is calculated by normalising each of the five statistics such that the individual IPE for each measure ranges from 0 for the best model to 1 for the worst. The coordinates of the ideal point using the five chosen measures are: RMSE = 0, RSqr = 1, ME = 0, PI = 1, PEP = 0. IPE, shown in Equation (24), measures how far a model is from this ideal point. In this equation i represents each of the models under

scrutiny (i.e. the 60 forecast cases). Note, it is possible to extend this equation to cover j different models across k different datasets, but this extension is not required here. Each IPE is specific to the study undertaken – in other words, one cannot use it to compare models across different studies, it can only be used to compare models within the same study:

$$\mathrm{IPE} = \left[0.2\left(\left(\frac{RMSE_i}{\max RMSE}\right)^2 + \left(\frac{RSqr_i - 1}{\min RSqr - 1}\right)^2 + \left(\frac{ME_i}{\max |ME|}\right)^2 + \left(\frac{PI_i - 1}{\min PI - 1}\right)^2 + \left(\frac{PEP_i}{\max |PEP|}\right)^2\right)\right]^{\frac{1}{2}} \quad (24)$$

IPE is made up, in this case, of the five selected statistics. It is possible to incrementally add these statistics to IPE to determine just how much additional information each statistic provides in differentiating between model performances. For example, using RMSE alone results in a weighted RMSE evaluation metric. Adding in RSqr then produces a weighted, combined RMSE, RSqr statistic and so on (with the component multiplier in the first case being set at 0.5, then 0.33, then 0.25 and finally 0.2).

Figure 5 shows the effect of adding additional statistics to the IPE measure for the three different time periods that our models cover (daily, weekly, ten-day period). Looking first at the daily models – each of the lines in this figure represents a different catchment modelled at a daily time period (20 in total). At a daily time step the models are all reasonably accurate. Therefore, the RMSE value for the models is very similar and it is difficult to pick out a 'good' model based on this statistic alone. The figure shows all of the daily models clustered with IPE1 scores (i.e. weighted RMSE values) between 0 and 0.1. Adding in the RSqr statistic to the IPE (resulting in IPE2) provides much more information, as shown in Figure 5 (daily forecast). IPE2 in this case is differentiating between the models much more effectively – as shown by the spread of values in the figure. Adding in ME provides more information again (IPE3) as does the addition of PI (IPE4). After PEP is added to IPE (resulting in IPE5 – i.e. the full IPE measure in Equation (24)) there is very little change in the performance measure for each of the sites. IPE has, therefore, stabilised with four combined statistics. The inclusion of PEP provides no strong additional information that can be used to differentiate between the models. This

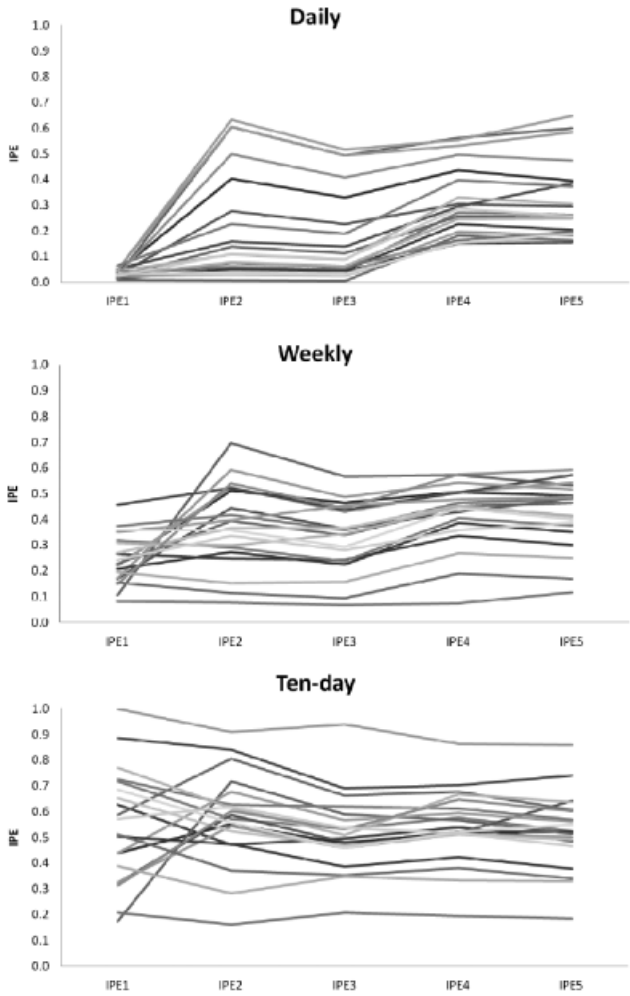

Figure 5 | Incremental IPE scores for models of the three different time periods.

is an alternative approach to determining an appropriate number of statistics to use following the PCA operation. In the case of daily models, in this study, only four statistics are needed to differentiate between each of the models' abilities.

Looking next at the weekly models – these models have a much more difficult task to perform than the daily models so it is not surprising that they perform somewhat differently to one another (some models are much more accurate than others). This is supported by the results shown in Figure 5 (weekly forecast). In this case the IPE1 scores are much more dispersed than the daily results, showing that the weighted RMSE values are very much different from one model to another. Adding the RSqr statistic (IPE2) results in even more

differentiation. In this case, having added the ME statistic (IPE3) the results stabilise and no further information is gained by adding PI or PEP. In this case, three statistics are sufficient to differentiate between the models. The ten-day models perhaps emphasise this point still further – again exhibiting stability when only three statistics are included in the IPE evaluation.

The results presented in this section show that only a small number of statistics need to be used to compare different models. In cases where models are similar to one another in terms of accuracy (for example, the daily models), more statistics are needed for comparison than when the models are dissimilar. In the case study presented here, four orthogonal statistics are useful for differentiating daily models, while three statistics are needed to differentiate weekly and ten-day models.

CONCLUSIONS

The use of an open access standard tool for assessing the performance of a forecast method permits a better understanding of the available hydrological modelling metrics. In this case, the availability of the HydroTest site has provided a fast performance assessment for 60 forecast cases from which it was possible to determine orthogonal composites and groups of redundant metrics. Being free of the performance assessment task permitted efforts to be focused on choosing a set of orthogonal metrics/criteria and to establish a general framework from which it was possible to define the relative forecasting capabilities of daily, weekly and ten-day water level models at 20 hydrological stations across a real-time hydrometerological network. The recommended use of a combined evaluation vector comprising three or four traditional metrics is the minimum required set of criteria that appears to be needed for analysis to establish the operational performance at a particular station in the network under test, depending on the forecast period. The number of output metrics that users needed to interpret was reduced from 22, to 5, to 1. The capability assessment presented here relates to the simple method that was implemented, for the purposes of initial testing, and it should also be remembered that a large number of the models did not fulfil the original calibration search procedure requirement. The use of additional predictor variables could improve the overall situation, providing enhanced evaluation results and a need for further testing.

ACKNOWLEDGEMENTS

The authors are grateful to the two anonymous referees who reviewed an earlier version of this paper for their constructive comments, which enhanced the quality and readability of the paper. This work was partially supported by Pontificia Universidad Javeriana (Research Project: 002185), Loughborough University and the University of Nottingham.

REFERENCES

Abrahart, R. J., Heppenstall, A. J. & See, L. M. 2007 Timing error correction procedure applied to neural network rainfall-runoff modelling. *Hydrol. Sci. J.* **52**(3), 414–431.

ASCE 1993 Criteria for evaluating watershed models. *J. Irrig. Drainage Engng.* **119**(3), 429–442.

Beven, K. 2002 Towards a coherent philosophy for modelling the environment. *Proc. R. Soc. Lond. A* **458**(2026), 2465–2484.

Costa, C., Rivera, H. & González, H. 2005 Aspectos conceptuales. In: *Protocolo para la emisión de pronósticos hidrológicos* Rivera, H. (Ed.), Imprenta Nacional, Bogotá, pp. 13–28.

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

Dawson, C. W., Abrahart, R. J. & See, L. M. 2010 HydroTest: further development of a web resource for the standardised assessment of hydrological models. *Environ. Modell. Software* **25**(11), 1481–1482.

Dawson, C. W., See, L. M., Abrahart, R. J. & Heppenstall, A. J. 2006 Symbiotic adaptive neuro-evolution applied to rainfall-runoff modelling in northern England. *Neural Net.* **19**(2), 236–247.

Domínguez, E. 1998 Protocolo de modelación matemática de procesos hidrológicos. *Meteorol. Colomb.* **2**, 33–38.

Domínguez, E. 2004 Aplicación de la ecuación de Fokker–Planck–Kolmogorov para el pronóstico de afluencias a embalses hidroeléctricos (caso práctico de la represa de Betania). *Meteorol. Colomb.* **8**, 17–26.

Domínguez, E. 2007a Introducción a la modelación matemática. Available at: http://www.mathmodelling.org.

Domínguez, E. 2007b *Pruebas piloto de modelación hidrológica para emitir pronósticos hidrológicos en forma cuantitativa y con uso de modelos auto-regresivos para el horizonte diario, pentadal*

y decadal de los niveles de agua – Viabilidad científica de la implementación de pronósticos operativos de los niveles del agua a escala diaria, pentadal y decadal. IDEAM Contrato 108-2007 Informe Final, Instituto de Hidrología, Meteorología y Estudios Ambientales - IDEAM, Bogotá.

Domínguez, E., Angarita, H., Ardila, F. & Caicedo, F. 2009 Hydrological risk modeling using adaptive operators: overview and applications. In: *Proc. 8th International Conference in Hydroinformatics*. IAHR, Concepción, Chile, p 12.

Domínguez, E., Angarita, H. & Rivera, H. 2010 The feasibility of daily, weekly and ten-day water-level forecasting in Colombia. *Ingeniería e Investigación*. **30**(2), 178–187.

Dunteman, G. H. 1989 *Principal Component Analysis*. Quantitative applications in the social sciences. Sage University Papers, London

Elshorbagy, A., Corzo, G., Srinivasulu, S. & Solomatine, D. 2010 Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part 1: concepts and methodology. *Hydrol. Earth Syst. Sci.***14**, 1931–1941.

Jolliffe, I. T. 1986 *Principal Component Analysis*. Springer, New York.

Kaiser, H. F. 1958 The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Wat. Res. Res.* **35**(1), 233–241.

Mussy, A. 2005 *Short Term Hydrological Forecasting Model In Colombia: Simulation For The Magdalena River*. Final Report, IDEAM, Lausanne

Napolitano, G., See, L., Calvo, B., Savi, F. & Heppenstall, A. 2010 A conceptual and neural network model for real-time flood forecasting of the Tiber River in Rome. *Physics and Chemistry of the Earth* **35**(3–5) 187–194.

Oreskes, N., Shrader-Frechette, K. & Belitz, K. 1994 Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **263**(5147), 641–646.

Pokhrel, P., Yilmaz, K. & Gupta, H. In Press Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *J. Hydrol*.

Refsgaard, J. C. & Henriksen, H. J. 2004 Modelling guidelines–terminology and guiding principles. *Adv. Wat. Res*. **27**(1), 71–82.

Rivera, H., Domínguez, E., González, H. & Zamudio, H. 2005 La modelación hidrológica específica para la emisión de pronósticos hidrológicos. In *Protocolo para la emisión de los pronósticos hidrológicos* (ed. in C. Costa), pp. 43–106. Imprenta Nacional, Bogotá.

Rivera, H., Zamudio, E. & Romero, H. 2004 Modelación con fines de pronósticos hidrológicos de los niveles diarios en periodo de estiaje en los sitios de Calamar, El Banco y Puerto Berrio del Magdalena. *Av. Recursos Hidráulicos* **11**, 115–130.

Rojas, N. 2006 *Determinación del flujo de la información hidrológica en tiempo real en los pronósticos hidrológicos del nivel del agua para la navegación del río magdalena*. Subdirección de Hidrología, IDEAM, Bogotá

Shaefli, B. & Gupta, H. V. 2007 Do Nash values have value? *Hydrol. Process.* **25**(15), 2075–2080.

Van Waveren, R. H., Groot, S., Scholten, S., Van Geer, F. C., Wösten, J. H. M., Koeze, R. D. & Noort, J. J. 1999 *Smooth Modelling in Water Management, Good Modelling Practice Handbook*. Dutch Dept. of Public Works, Institute for Inland Water Management and Waste Water Treatment, Lelystad.

Weglarczyk, S. 1998 The interdependence and applicability of some statistical quality measures for hydrological models. *J. Hydrol.* **206**(1), 98–103.

Willems, P. 2009 A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. *Environ. Modell. Software* **24**(3), 311–321.

First received 20 March 2009; accepted in revised form 21 March 2010. Available online 22 December 2010