


On the criteria of model performance evaluation for real-time flood forecasting

Ke-Sheng Cheng^{1,2}  · Yi-Ting Lien³ · Yii-Chen Wu¹ · Yuan-Fong Su⁴

Published online: 4 October 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Model performance evaluation for real-time flood forecasting has been conducted using various criteria. Although the coefficient of efficiency (*CE*) is most widely used, we demonstrate that a model achieving good model efficiency may actually be inferior to the naïve (or persistence) forecasting, if the flow series has a high lag-1 autocorrelation coefficient. We derived sample-dependent and AR model-dependent asymptotic relationships between the coefficient of efficiency and the coefficient of persistence (*CP*) which form the basis of a proposed *CE–CP* coupled model performance evaluation criterion. Considering the flow persistence and the model simplicity, the AR(2) model is suggested to be the benchmark model for performance evaluation of real-time flood forecasting models. We emphasize that performance evaluation of flood forecasting models using the proposed *CE–CP* coupled criterion should be carried out with respect to individual flood events. A single *CE* or *CP* value derived from a multi-event artificial series by no means provides a multi-event overall evaluation and may actually disguise the real capability of the proposed model.

Keywords Model performance evaluation · Uncertainty · Coefficient of persistence · Coefficient of efficiency · Real-time flood forecasting · Bootstrap

1 Introduction

Like many other natural processes, the rainfall–runoff process is composed of many sub-processes which involve complicated and scale-dependent temporal and spatial variations. It appears that even less complicated hydrological processes cannot be fully characterized using only physical models, and thus many conceptual models and physical models coupled with random components have been proposed for rainfall–runoff modeling (Nash and Sutcliffe 1970; Bergström and Forsman 1973; Bergström 1976; Rodríguez-Iturbe and Valdés 1979; Rodríguez-Iturbe et al. 1982; Lindström et al. 1997; Du et al. 2009). These models are established based on our understanding or conceptual perception about the mechanisms of the rainfall–runoff process.

In addition to pure physical and conceptual models, empirical data-driven models such as the artificial neural networks (ANN) models for runoff estimation or forecasting have also gained much attention in recent years. These models usually require long historical records and lack physical basis. As a result, they are not applicable for ungauged watersheds (ASCE 2000). The success of an ANN application depends both on the quality and the quantity of the available data. This requirement cannot be easily met, as many hydrologic records do not go back far enough (ASCE 2000).

Almost all models need to be calibrated using observed data. This task encounters a range of uncertainties which stem from different sources including data uncertainty,

✉ Ke-Sheng Cheng
rslab@ntu.edu.tw

¹ Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan, ROC

² Master Program in Statistics, National Taiwan University, Taipei, Taiwan, ROC

³ TechNews, Inc., Taipei, Taiwan, ROC

⁴ National Science and Technology Center for Disaster Reduction, Taipei, Taiwan, ROC

parameter uncertainty, and model structure uncertainty (Wagener et al. 2004). The uncertainties involved in model calibration will unavoidably propagate to the model outputs. The simple regression models and ANN models are strongly dependent on the data used for calibration and their reliability beyond the range of observations may be questionable (Michaud and Sorooshian 1994; Refsgaard 1994). Researchers have also found that many hydrological processes are complicated enough to allow for different parameter combinations (or parameter sets), often widely distributed over their individual feasible ranges, to yield similar or compatible model performances (Beven 1989; Kuczera 1997; Kuczera and Mroczkowski 1998; Wagener et al. 2004; Wagener and Gupta 2005). This is known as the problem of parameter or model identifiability, and the effect is referred to as parameter or model equifinality (Beven and Binley 1992; Beven 1993, 2006). A good discussion about the parameter or model equifinality was given by Lee et al. (2012).

Since the uncertainties in model calibration can be propagated to the model outputs, performance of hydrological models must be evaluated considering the uncertainties in model outputs. This is usually done by using another independent set of historical or observed data and employing different evaluation criteria. A few criteria have been adopted for model performance evaluation (hereinafter abbreviated as MPE), including the root-mean-squared error (*RMSE*), correlation coefficient, coefficient of efficiency (*CE*), coefficient of persistence (*CP*), peak error in percentages (E_{Op}), mean absolute error (*MAE*), etc. The concept of choosing benchmark series as the basis for model performance evaluation was proposed by Seibert (2001). Different criteria evaluate different aspects of the model performance, and using a single criterion may not always be appropriate. Seibert and McDonnell (2002) demonstrated that simply modeling runoff with a high coefficient of efficiency is not a robust test of model performance. Due to the uncertainties in the model outputs, a specific MPE criterion can yield a range of different values which characterizes the uncertainties in model performance. A task committee of the American Society of Civil Engineers (ASCE 1993) conducted a thorough review on criteria for models evaluation and concluded that—“There is a great need to define the criteria for evaluation of watershed models clearly so that potential users have a basis with which they can select the model best suited to their needs”.

The objectives of this study are three-folds. Firstly, we aim to demonstrate the effects of parameter and model structure uncertainties on the uncertainty of model outputs through stochastic simulation of exemplar hydrological processes. Secondly, we intend to evaluate the effectiveness of different criteria for model performance evaluation.

Lastly, we aim to investigate the theoretical relationship between two MPE criteria, namely the coefficient of efficiency and coefficient of persistence, and to propose a *CE–CP* coupled criteria for model performance evaluation. In this study we focus our analyses and discussions on the issue of real-time flood forecasting.

The remainder of this paper is organized as follows. Section 2 describes some natures of flood flow forecasting that should be considered in evaluating model performance evaluation. In Sect. 3, we introduce some commonly used criteria for model performance evaluation and discuss their properties. In Sect. 4, we demonstrate the parameter and model uncertainties and uncertainties in criteria for model performance evaluation by using simulated AR series. Section 5 gives a detailed derivation of an asymptotic sample-dependent *CE–CP* relationship which is used to determine whether a forecasting model with a specific *CE* value can be considered as achieving better performance than the naïve forecasting. Section 6 introduces the idea of using the AR(2) model as the benchmark for model performance evaluation and derives the model-dependent *CE–CP* relationships for AR(1) and AR(2) models. These relationships form the basis for a *CE–CP* coupled approach of model performance evaluation. In Sect. 7, the *CE–CP* coupled approach to model performance evaluation was implemented using bootstrap samples of historical flood events. Discussions on calculation of *CE* values for multi-event artificial series and single-event series are also given in Sect. 7. Section 8 discusses usage of *CP* for performance evaluation of multiple-step forecasting. Section 9 gives a summary and concluding remarks of this study.

2 Some natures of flow forecasting

A hydrological process often consists of many sub-processes which cannot be fully characterized by physical laws. For some applications, we are not even sure whether all sub-processes have been considered. The lack of full knowledge of the hydrological process under investigation inevitably leads to uncertainties in model parameters and model structure when historical data are used for model calibration.

Another important issue which is critical to hydrological forecasting is our limited capability of observing hydrological variables in a spatiotemporal domain. Hydrological processes occur over a vast spatial extent and it is usually impossible to observe the process with adequate spatial density and resolution over the entire study area. In addition, temporal variations of hydrological variables are difficult to be described solely by physical governing equations, and thus stochastic components need to be incorporated or stochastic models be developed to

characterize such temporal variations. Due to our inability of observing and modeling the spatiotemporal variations of hydrological variables, performance of flood forecasting models can vary from one event to another, and stochastic models are sought after for real-time flood forecasting. In recent years, flood forecasting models that incorporating ensemble of numerical weather predictions derived from weather radar or satellite observations have also gained great attention (Cloke and Pappenberger 2009). Flood forecasting systems that integrate rainfall monitoring and forecasting with flood forecasting and warning are now operational in many areas (Moore et al. 2005).

The target variable or the model output of a flood forecasting model is the flow or the stage at the watershed outlet. A unique and important feature of the flow at the watershed outlet is its temporal persistence. Even though the model input (rainfalls) may exhibit significant spatial and temporal variations, flow at the watershed outlet is generally more persistent in time. This is due to the buffering effect of the watershed which helps to dampen down the effect of spatial and temporal variations of rainfalls on temporal variation of flow at the outlet. Such flow persistence indicates that previous flow observations

can provide valuable information for real-time flow forecasting.

If we consider the flow time series as the following stationary autoregressive process of order p (AR(p)),

$$x_t = \phi_0 + \sum_{i=1}^p \phi_i x_{t-i} + \varepsilon_t \tag{1}$$

where x_t and ε_t respectively represent the flow and noise at time t , and ϕ_i 's are parameters of the model. A measure of persistence can then be defined as the *cumulative impulse response* (CIR) of the AR(p) process (Andrews and Chen 1994), i.e.,

$$CIR = \frac{1}{1 - \rho}, \tag{2}$$

$$\rho = \sum_{i=1}^p \phi_i. \tag{3}$$

Figure 1 demonstrates the persistence feature of flows at the watershed outlet. The watershed (Chi-Lan River watershed in southern Taiwan) has a drainage area of approximately 110 km² and river length of 19.16 km. Partial autocorrelation functions of the rainfall and flow

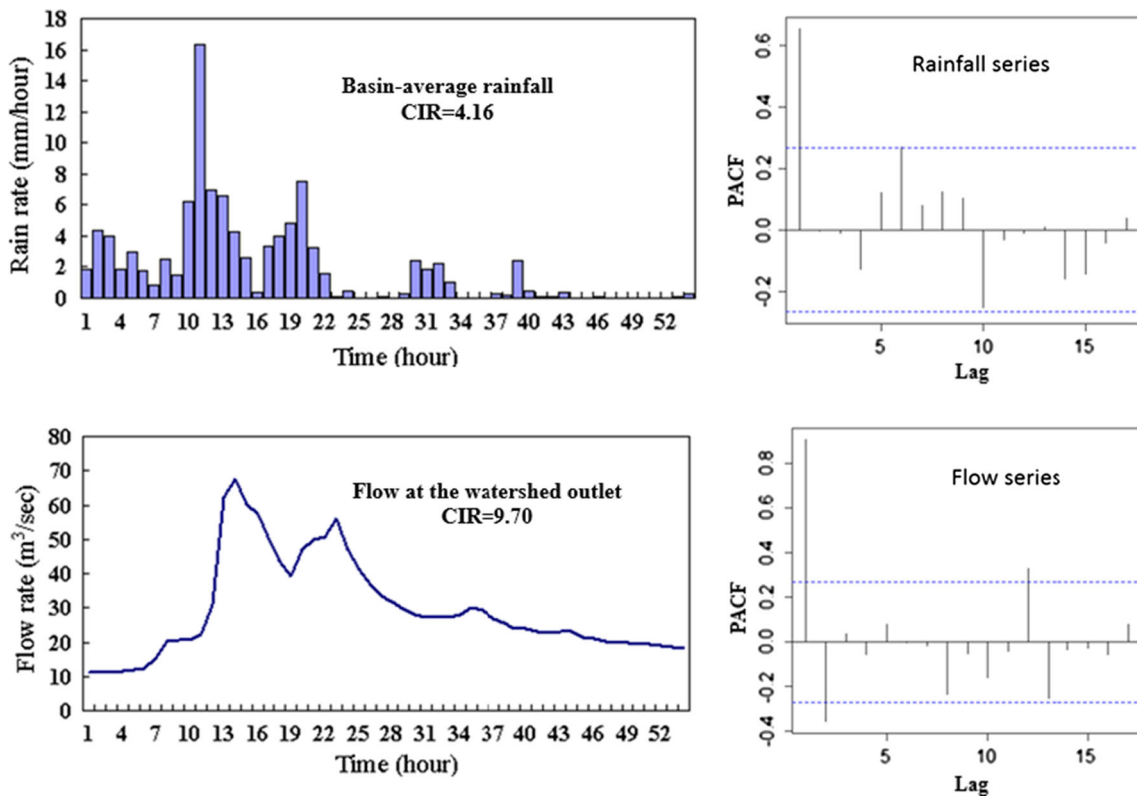


Fig. 1 An example showing higher persistence for flow at the watershed outlet than the basin-average rainfall. The cumulative impulse response (CIR) represents a measure of persistence (CIR). The partial autocorrelation functions (PACF) of the rainfall and flow

series are also shown. Dashed lines in the PACF plots represent the upper and lower limits of the critical region, at a 5 % significance level, of a test that a given partial correlation is zero

series (see Fig. 1) show that for the rainfall series, only the lag-1 partial autocorrelation coefficient is significantly different from zero, whereas for the flow series, the lag-1 and lag-2 partial autocorrelation coefficients are significantly different from zero. Thus, basin-average rainfalls of the event in Fig. 1 was modeled as an AR(1) series and flows at the watershed outlet were modeled as an AR(2) series. CIR values of the rainfall series and the flow series are 4.16 and 9.70, respectively. The flow series have significantly higher persistence than the rainfall series. We have analyzed flow data at other locations and found similar high persistence in flow data series.

3 Criteria for model performance evaluation

Evaluation of model performance can be conducted by graphical or quantitative methods. The former graphically compares time series plots of the predicted series and the observed series, whereas the latter uses numerical indices as evaluation criteria. Figures intended to show how well predictions agree with observations often only provide limited information because long series of predicted data are squeezed in and lines for observed and predicted data are not easily distinguishable. Such evaluation is particularly questionable in cases that several independent events were artificially combined to form a long series of predicted and observed data. Lagged-forecasts could have occurred in individual events whereas the long artificial series still appeared to provide perfect forecasts in such squeezed graphical representations. Not all authors provide numerical information, but only state that the model was in ‘good agreement’ with the observations (Seibert 1999). Thus, in addition to graphical comparison, model performance evaluation using numerical criteria is also desired.

While quite a few MPE criteria have been proposed, researchers have not had consensus on how to choose the best criteria or what criteria should be included at the least. There are also cases of *ad hoc* selection of evaluation criteria in which the same researchers may choose different criteria in different study areas for applications of similar natures. Table 1 lists criteria used by different applications. Definitions of these criteria are given as follows.

- (1) Relative error (*RE*)

$$RE_t = \frac{|Q_t - \hat{Q}_t|}{Q_t} \times 100\% \tag{4}$$

Q_t is the observed data (Q) at time t , \hat{Q}_t is the predicted value at time t .

The relative error is used to identify the percentage of samples belonging to one of the three groups:

“low relative error” with $RE \leq 15\%$, “medium error” with $15\% < RE \leq 35\%$, and “high error” with $RE > 35\%$ (Corzo and Solomatine 2007).

- (2) Mean absolute error (*MAE*)

$$MAE = \frac{1}{n} \sum_{t=1}^n |Q_t - \hat{Q}_t| \tag{5}$$

n is the number of data points.

- (3) Correlation coefficient (*r*)

$$r = \frac{\sum_{t=1}^n (Q_t - \bar{Q})(\hat{Q}_t - \bar{\hat{Q}})}{\sqrt{\sum_{t=1}^n (Q_t - \bar{Q})^2} \sqrt{\sum_{t=1}^n (\hat{Q}_t - \bar{\hat{Q}})^2}} \tag{6}$$

\bar{Q} is the mean of observed Q , $\bar{\hat{Q}}$ is the mean of predicted flow \hat{Q} .

- (4) Root-mean-squared error (*RMSE*)

$$RMSE = \sqrt{\frac{SSE}{n}}, \tag{7a}$$

$$SSE = \sum_{t=1}^n (Q_t - \hat{Q}_t)^2 \tag{7b}$$

- (5) Normalized root-mean-squared error (*NRMSE*) (Corzo and Solomatine 2007; Pebesma et al. 2007)

$$NRMSE = \frac{RMSE}{s_{obs}} \tag{8a}$$

s_{obs} is the sample standard deviation of observed data Q . or

$$NRMSE = \frac{RMSE}{\bar{Q}} \tag{8b}$$

- (6) Coefficient of efficiency (*CE*) (Nash and Sutcliffe 1970)

$$CE = 1 - \frac{SSE}{SST_m} = 1 - \frac{\sum_{t=1}^n (Q_t - \hat{Q}_t)^2}{\sum_{t=1}^n (Q_t - \bar{Q})^2} \tag{9}$$

\bar{Q} is the mean of observed data Q . SST_m is the sum of squared errors with respect to the mean value.

- (7) Coefficient of persistence (*CP*) (Kitanidis and Bras 1980)

$$CP = 1 - \frac{SSE}{SSE_N} = 1 - \frac{\sum_{t=1}^n (Q_t - \hat{Q}_t)^2}{\sum_{t=1}^n (Q_t - Q_{t-k})^2} \tag{10}$$

SSE_N is the sum of squared errors of the naïve (or persistent) forecasting model ($\hat{Q}_t = Q_{t-k}$)

- (8) Error in peak flow (or stage) in percentages or absolute value (*Ep*)

Table 1 Summary of criteria for model performance evaluation

Applications	Criteria								Target variable
	RMSE	r ^a	CE	CP	MAE	NRMSE	Ep	RE	
Schreider et al. (1997)			✓	✓	✓				Flow
Labat et al. (1999)			✓						Flow
Yu et al. (2000)			✓						Flow
Markus et al. (2003)	✓				✓				Water quality
Ancil and Rat (2005)				✓					Flow
Sarangi and Bhattacharya (2005)		✓	✓		✓				Sediment yield
Lauzon et al. (2006)	✓			✓					Flow
Sahoo et al. (2006)	✓	✓			✓				Flow, water quality
Corzo and Solomatine (2007)			✓	✓		✓		✓	Flow
Coulibaly and Evora (2007)		✓			✓				Precipitation, temperature
Dibike and Coulibaly (2007)	✓		✓						Flow
Harmel and Smith (2007)	✓		✓		✓				Flow, water quality
Pebesma et al. (2007)	✓	✓	✓			✓			Flow
Calvo and Savi (2009)				✓			✓		Flow
Chang et al. (2009)	✓			✓	✓				Flow
Lin et al. (2009)	✓	✓	✓		✓				Flow
Sauter et al. (2009)	✓	✓	✓						Water level
Wang et al. (2010)	✓		✓				✓		Flow
Wu et al. (2010)	✓		✓	✓					Rainfall
Sattari et al. (2012)	✓	✓			✓	✓			Flow
Chen et al. (2013)	✓	✓	✓	✓	✓				Flow
Kasiviswanathan and Sudheer (2013)			✓						Flow
Chiew et al. (2014)			✓						Flow
Wang et al. (2014)			✓						Flow
Counts of applications	13	8	16	8	10	3	2	1	

^a Including applications using coefficient of determination (*r*²)

$$Ep = \frac{Q_p - \hat{Q}_p}{Q_p} \times 100\% \tag{11}$$

Q_p is the observed peak value, *Q̂_p* is the predicted peak value.

From Table 1, we found that *RMSE*, *CE* and *MAE* were most widely used, and, except for Yu et al. (2000), all applications used multi-criteria for model performance evaluation.

Generally speaking, model performance evaluation aims to assess the goodness-of-fit of the model output series to the observed data series. Thus, except for *Ep* which is a local measure, all other criteria can be viewed as goodness-of-fit measures. The *CE* evaluates the model performance with reference to the mean of the observed data. Its value can vary from 1, when there is a perfect fit, to $-\infty$. A negative *CE* value indicates that the model predictions are worse than predictions using a constant equal to the average of the observed data. For linear

regression models, *CE* is equivalent to the coefficient of determination *r*². It has been found that *CE* is a much superior measure of goodness-of-fit compared with the coefficient of determination (Willmott 1981; Legates and McCabe 1999; Harmel and Smith 2007). Moriasi et al. (2007) recommended the following model performance ratings:

- CE* ≤ 0.50 unsatisfactory
- 0.50 < *CE* ≤ 0.65 satisfactory
- 0.50 < *CE* ≤ 0.65 good
- 0.75 < *CE* ≤ 1.00 very good

However, Moussa (2010) demonstrated that good simulations characterized by *CE* close to 1 can become “monsters” if other model performance measures (such as *CP*) had low or even negative values.

Although not widely used for model performance evaluation, usage of the coefficient of persistence was also advocated by some researchers (Kitanidis and Bras 1980; Gupta et al. 1999; Lauzon et al. 2006; Corzo and

Solomatine 2007; Calvo and Savi 2009; Wu et al. 2010). The coefficient of persistence is a measure that compares the performance of the model being used and performance of the naïve (or persistent) model which assumes a steady state over the forecast lead time. Equation (10) represents the CP of a k -step lead time forecasting model since Q_{t-k} is used in the denominator. The CP can assume a value between $-\infty$ and 1 which indicates a perfect model performance. A small positive value of CP may imply occurrence of lagged prediction, whereas a negative CP value indicates that performance of the model being used is inferior to the naïve model. Gupta et al. (1999) indicated that the coefficient of persistence is a more powerful test of model performance (i.e. capable of clearly indicating poor model performance) than the coefficient of efficiency.

Standard practice of model performance evaluation is to calculate CE (or some other common performance measure) for both the model and the naïve forecast, and the model is only considered acceptable if it beats persistence. However, from the research works listed in Table 1, most research works which conducted model performance evaluation did not pay much attention to whether the model performed better than a naïve persistence forecast. Yaseen et al. (2015) also explored comprehensively the literature on the applications of artificial intelligent for flood forecasting. Their survey revealed that the coefficient of persistence was not widely adopted for model performance evaluation. Moriasi et al. (2007) also reported that the coefficient of persistence has been used only occasionally in the literature, so a range of reported values is not available.

Calculations of CE and CP differ only in the denominators which specify what the predicted series are compared against. Seibert (2001) addressed the importance of choosing an appropriate *benchmark series* which forms the basis for model performance evaluation. The following *bench coefficient* (G_{bench}) can be used to compare the goodness-of-fit of the predicted series and the benchmark series to the observed data series (Seibert 2001).

$$G_{bench} = 1 - \frac{\sum_{t=1}^n (Q_t - \hat{Q}_t)^2}{\sum_{t=1}^n (Q_t - Q_{b,t})^2}, \quad (12)$$

$Q_{b,t}$ is the value of the benchmark series Q_b at time t .

The bench coefficient provides a general form for measures of goodness-of-fit based on benchmark comparisons. The CE and CP are bench coefficients with respect to benchmark series of the constant mean and the naïve-forecast, respectively. The bottom line, however, is what benchmark series should be used for the target application.

4 Model performance evaluation using simulated series

As we have mentioned in Sect. 2, flows at the watershed outlet exhibit significant persistence and time series of streamflows can be represented by an autoregressive model. In addition, a few studies have also demonstrated that, with real-time error correction, AR(1) and AR(2) can significantly enhance the reliability of the forecasted water stages at the 1-, 2-, and 3-h lead time (Wu et al. 2012; Shen et al. 2015). Thus, we suggest using the AR(2) model as the benchmark series for flood forecasting model performance evaluation. In this section we demonstrate the parameter and model structure uncertainties using random samples of AR(2) models.

4.1 Parameter and model structure uncertainties

In order to demonstrate uncertainties involved in model calibration and to assess the effects of the parameter and model structure uncertainties on MPE criteria, sample series of the following AR(2) model were generated by stochastic simulation

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t, \quad (13)$$

$$\varepsilon_t \sim iid \quad N(0, \sigma_\varepsilon^2), \quad |\phi_2| < 1, \quad -1 < \frac{\phi_1}{1 - \phi_2} < 1$$

It can be shown that the AR(2) model has the following properties:

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}, \quad (14)$$

$$\rho_2 = \frac{\phi_1^2}{1 - \phi_2} + \phi_2, \quad (15)$$

and

$$\sigma_X^2 = \frac{\sigma_\varepsilon^2}{(1 - \phi_1 \rho_1 - \phi_2 \rho_2)} \quad (16)$$

where ρ_1 and ρ_2 are respectively lag-1, lag-2 autocorrelation coefficients of the random process $\{X_t, t = 1, 2, \dots\}$, and σ_X^2 is the variance of the random variable X .

For our simulation, parameters ϕ_1 and ϕ_2 were set to be 0.5 and 0.3 respectively, while four different values (1, 3, 5, and 7) were set for the parameter σ_ε . Such parameter setting corresponds to values of 1.50, 4.49, 7.49, and 10.49 for the standard deviation of the random variable X . For each $(\phi_1, \phi_2, \sigma_\varepsilon)$ parameter set, 1000 sample series were generated. Each series is composed of 1000 data points and is expressed as $\{x_i, i = 1, 2, \dots, 1000\}$. We then divided each series into a *calibration subseries* including the first

800 data points and a *forecast subseries* consisting of the remaining 200 data points. Parameters ϕ_1 and ϕ_2 were then estimated using the calibration subseries $\{x_i, i = 1, \dots, 800\}$. These parameter estimates ($\hat{\phi}_1$ and $\hat{\phi}_2$) were then used for forecasting with respect to the forecast subseries $\{x_i, i = 801, \dots, 1000\}$. In this study, only forecasting with one-step lead time was conducted. MPE criteria of *RMSE*, *CE* and *CP* were then calculated using simulated subseries $\{x_i, i = 801, \dots, 1000\}$ and forecasted subseries $\{\hat{x}_i, i = 801, \dots, 1000\}$. Each of the 1000 sample series was associated with a set of MPE criteria (*RMSE*, *CE*, *CP*), and uncertainty assessment of the MPE criteria was conducted using these 1000 sets of (*RMSE*, *CE*, *CP*). The above process is illustrated in Fig. 2.

Histograms of parameter estimates ($\hat{\phi}_1, \hat{\phi}_2$) with respect to different values of σ_ε are shown in Fig. 3. Averages of parameter estimates are very close to the theoretical value ($\phi_1 = 0.5, \phi_2 = 0.3$) due to the asymptotic unbiasedness of the maximum likelihood estimators. Uncertainties in parameter estimation are characterized by the standard deviation of $\hat{\phi}_1$ and $\hat{\phi}_2$. Regardless of changes in σ_ε , parameter uncertainties, i.e. $s_{\hat{\phi}_1}$ and $s_{\hat{\phi}_2}$, remain nearly constant, indicating that parameter uncertainties only depend on the length of the data series used for parameter estimation. The maximum likelihood estimators $\hat{\phi}_1$ and $\hat{\phi}_2$

are correlated and can be characterized by a bivariate normal distribution, as demonstrated in Fig. 4. Despite changes in σ_ε , these ellipses are nearly identical, reasserting that parameter uncertainties are independent of the noise variance σ_ε^2 .

The above parameter estimation and assessment of uncertainties only involve parameter uncertainties, but not the model structure uncertainties since the sample series were modeled with a correct form. In order to assess the effect of model structure uncertainties, the same sample series were modeled by an AR(1) model through a similar process of Fig. 2. Histogram of AR(1) parameter estimates ($\hat{\phi}_1$) with respect to different values of σ_ε are shown in Fig. 5. Averages of $\hat{\phi}_1$ with respect to various values of σ_ε are approximately 0.71 which is significantly different from the AR(2) model parameters ($\phi_1 = 0.5, \phi_2 = 0.3$) owing to the model specification error. Parameter uncertainties ($s_{\hat{\phi}_1}$) of AR(1) modeling, which are about the same magnitude as that of AR(2) modeling, are independent of the noise variance. It shows that the AR(1) model specification error does not affect the parameter uncertainties. However, the bias in parameter estimation of AR(1) modeling will result in a poorer forecasting performance and higher uncertainties in MPE criteria, as described in the next subsection.

Fig. 2 Illustrative diagram showing the process of (1) parameter estimation, (2) forecasting, (3) MPE criteria calculation, and (4) uncertainty assessment of MPE criteria

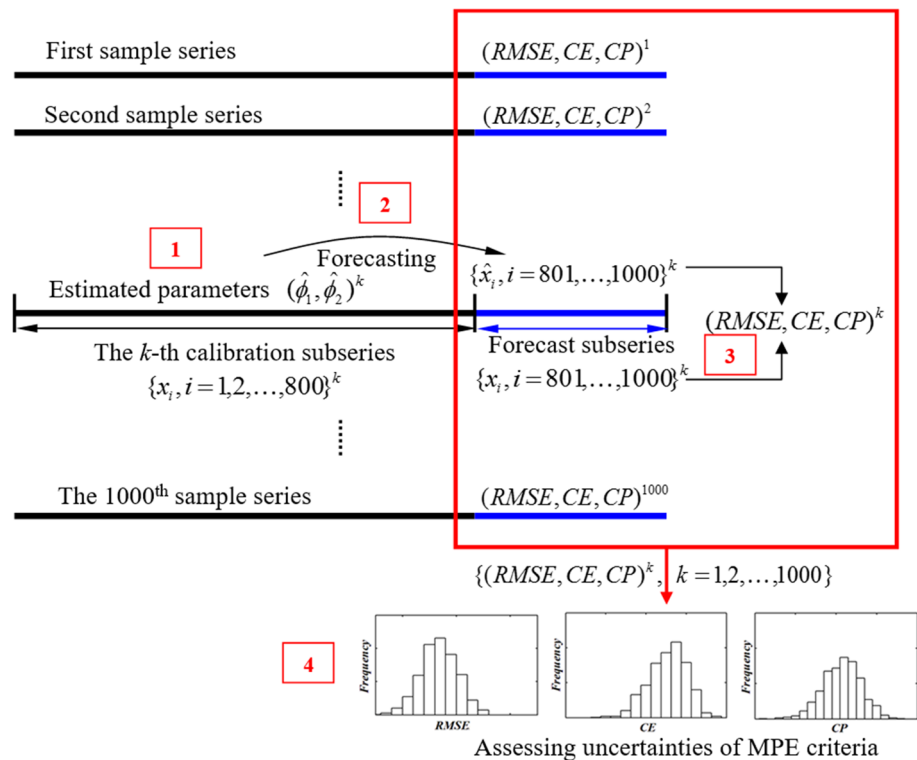
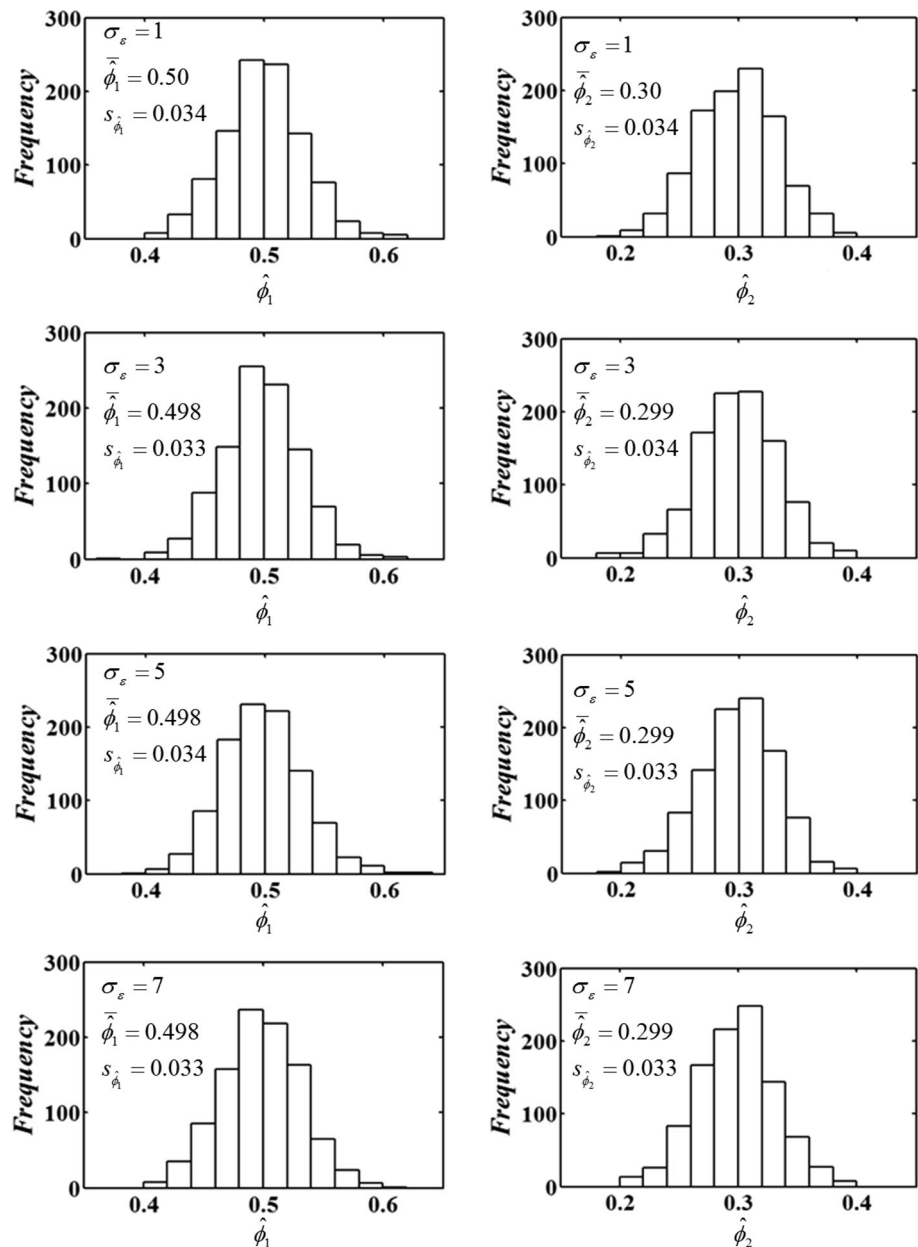


Fig. 3 Histograms of parameter estimates ($\hat{\phi}_1, \hat{\phi}_2$) using AR(2) model. Uncertainty in parameter estimation is independent of the noise variance σ_ε^2 . [Theoretical data model $X_t = 0.5X_{t-1} + 0.3X_{t-2} + \varepsilon_t$]



4.2 Uncertainties in MPE criteria

Through the process of Fig. 2, uncertainties in MPE criteria ($RMSE$, CE and CP) by AR(1) and AR(2) modeling and forecasting of the data series can be assessed. The $RMSE$ is dependent on σ_X which in turn depends on σ_ε . Thus, we evaluate uncertainties of the root-mean-squared errors normalized by the sample standard deviation s_X , i.e. $NRMSE$ (Eq. 8a). Figure 6 demonstrates the uncertainties of $NRMSE$ for the AR(1) and AR(2) modeling. AR(1) modeling of the sample series involves parameter uncertainties and model

structure uncertainties, while AR(2) modeling involves only parameter uncertainties. Although the model specification error does not affect parameter uncertainties, it results in bias in parameter estimation, and thus increases the magnitude of $NRMSE$. Mean value of $NRMSE$ by AR(2) modeling is about 95 % of the mean $NRMSE$ by AR(1) modeling. Standard deviation of $NRMSE$ by AR(2) modeling is approximately 88 % of the standard deviation of $NRMSE$ by AR(1) modeling. Such results indicate that presence of the model specification error results in a poorer performance with higher mean and standard deviation of $NRMSE$.

Fig. 4 Scatter plots of $(\hat{\phi}_1, \hat{\phi}_2)$ for AR(2) model with different values of σ_ε . Ellipses represent the 95 % density contours, assuming bivariate normal distribution for $\hat{\phi}_1$ and $\hat{\phi}_2$. [Theoretical data model $X_t = 0.5X_{t-1} + 0.3X_{t-2} + \varepsilon_t$]

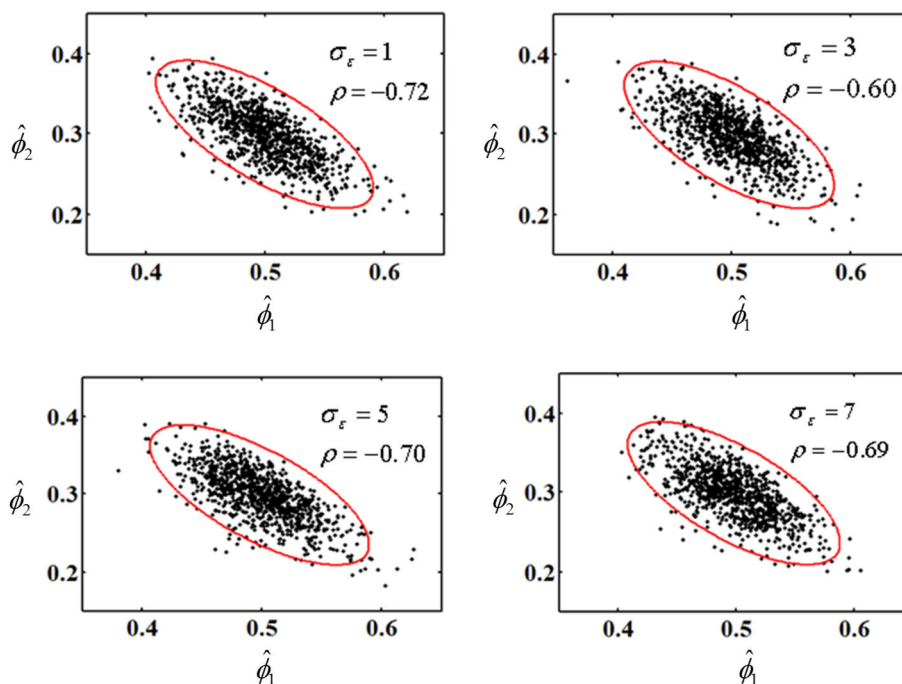
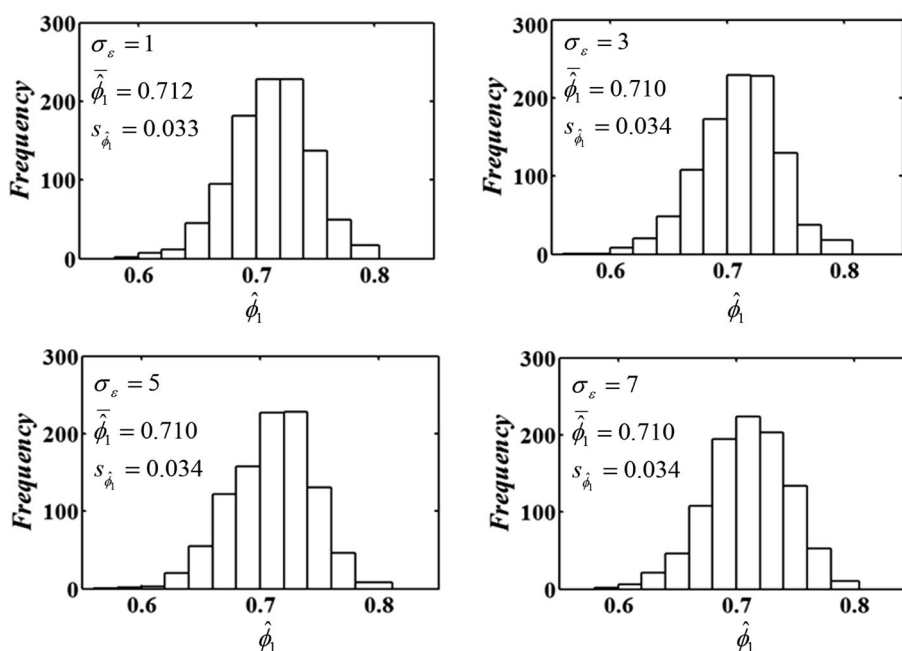


Fig. 5 Histograms of parameter estimates $(\hat{\phi}_1)$ using AR(1) model. Uncertainty in parameter estimation is independent of the noise variance σ_ε^2 . [Theoretical data model $X_t = 0.5X_{t-1} + 0.3X_{t-2} + \varepsilon_t$]

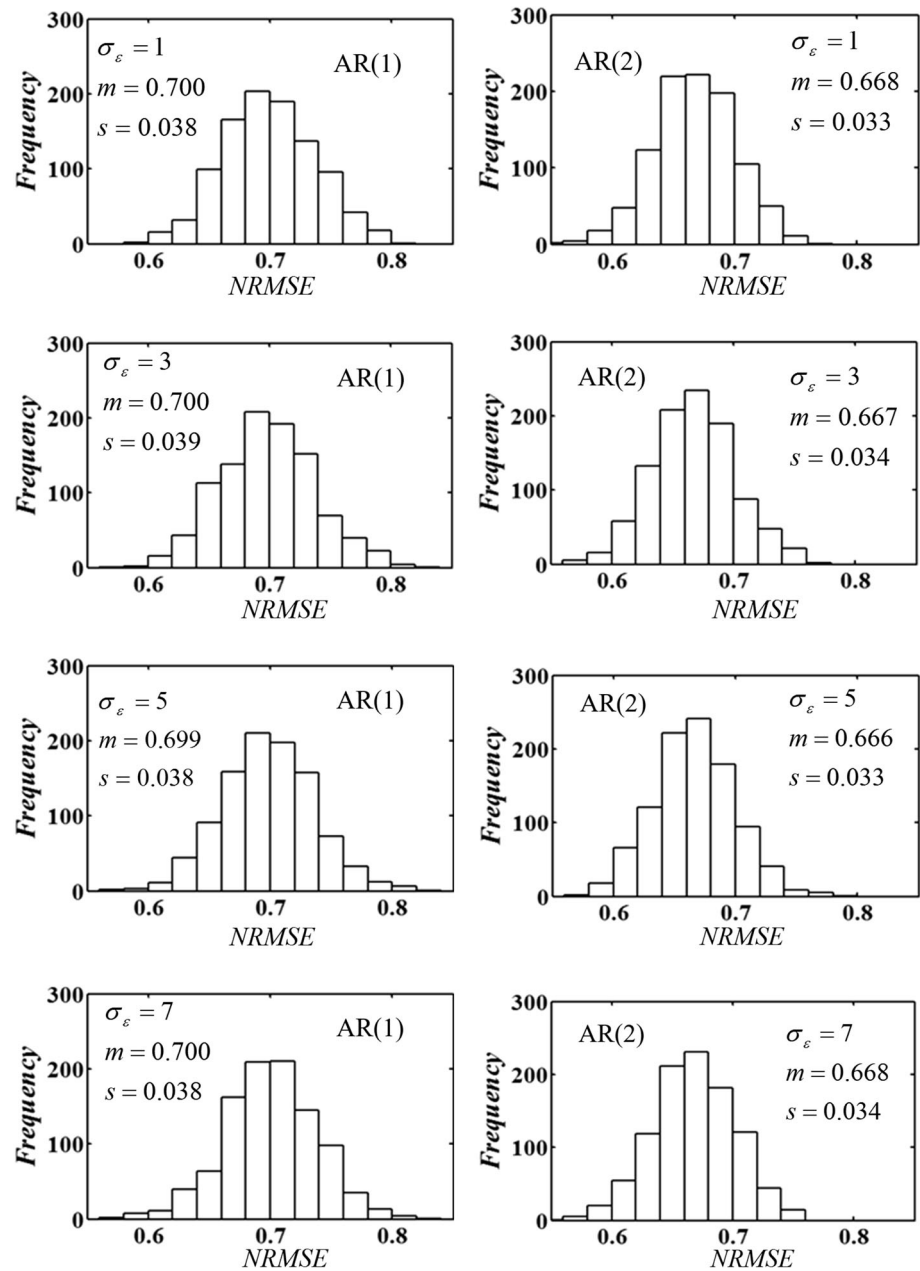


Histograms of *CE* and *CP* for AR(1) and AR(2) modeling of the data series are shown in Figs. 7 and 8, respectively. On average, *CE* of AR(2) modeling (without model structure uncertainties) is about 10 % higher than *CE* of AR(1) modeling. In contrast, the average *CP* of AR(2) modeling is approximately 55 % higher than the average *CP* of AR(1) modeling. The difference (measured in percentage) in the mean *CP* values of AR(1) and AR(2) modeling is larger than that of *CE* and *NRMSE*, suggesting that, for our exemplar AR(2) model, *CP* is a more sensitive

MPE criterion with presence of model structure uncertainty. Such results are consistent with the claim by Gupta et al. (1999) that the coefficient of persistence is a more powerful test of model performance. The reason for such results will be explained in the following section using an asymptotic relationship between *CE* and *CP*.

It is emphasized that we do not intend to mean that more complex models are not needed, but just emphasize that complex models may not always perform better than simpler models because of the possible

Fig. 6 Histograms of the normalized *RMSE* for AR(1) and AR(2) modeling with respect to various noise variance σ_ε^2

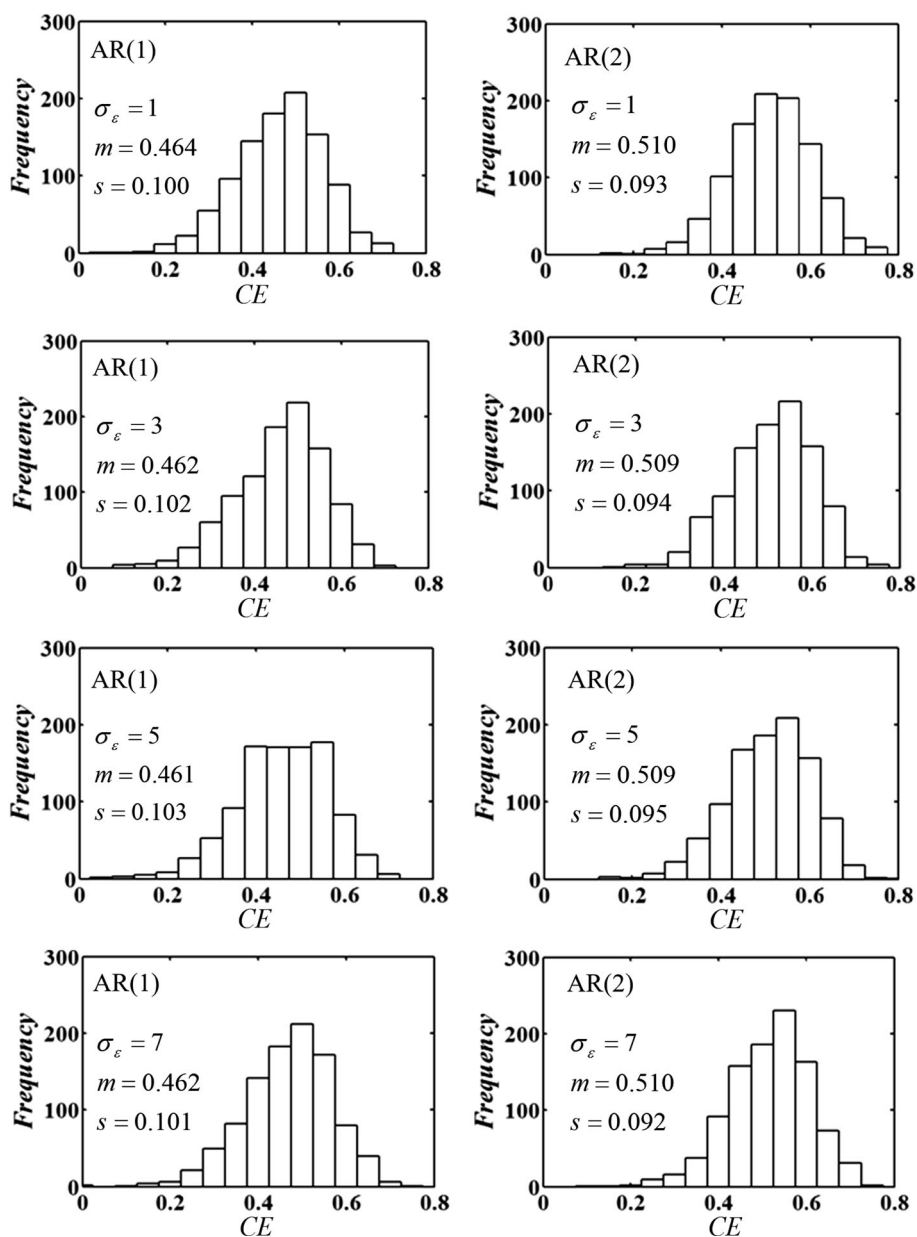


“over-parameterization” (Sivakumar 2008a). It is of great importance to identify the dominant processes that govern hydrologic responses in a given system and adopt practices that consider both simplification and generalization of hydrologic models (Sivakumar 2008b). Studies have also found that AR models were quite competitive with the complex nonlinear models including k -nearest neighbor and ANN models. (Tongal and Berndtsson 2016). In this regard, the significant flow persistence represents an important feature in flood forecasting and the AR(2) model is simple enough, while capturing the flow persistence, to suffice a bench mark series.

5 Sample-dependent asymptotic relationship between *CE* and *CP*

Given a sample series $\{x_t, t = 1, 2, \dots, n\}$ of a stationary time series, *CE* and *CP* respectively represent measures of model performance by choosing the constant mean series and the naive forecast series as the benchmark series. There exists an asymptotic relationship between *CE* and *CP* which should be considered when using *CE* alone for model performance evaluation. From the definitions of SST_m and SSE_N in Eqs. 9 and 10, for a k -step lead time forecast we have

Fig. 7 Histograms of the coefficient of efficiency (CE) for AR(1) and AR(2) modeling with respect to various noise variance σ_ε^2



$$\frac{SST_m}{n} \xrightarrow{n \rightarrow \infty} \sigma_X^2,$$

$$\frac{SSE_N}{n} \xrightarrow{n \rightarrow \infty} 2\sigma_X^2(1 - \rho_k).$$

And thus,

$$\frac{SST_m}{SSE_N} \xrightarrow{n \rightarrow \infty} \frac{1}{2(1 - \rho_k)}.$$

Therefore, for forecasting with a k -step lead time,

$$(17) \quad CE = 1 - \frac{SSE}{SST_m} = 1 - 2(1 - \rho_k) \frac{SSE}{SSE_N}$$

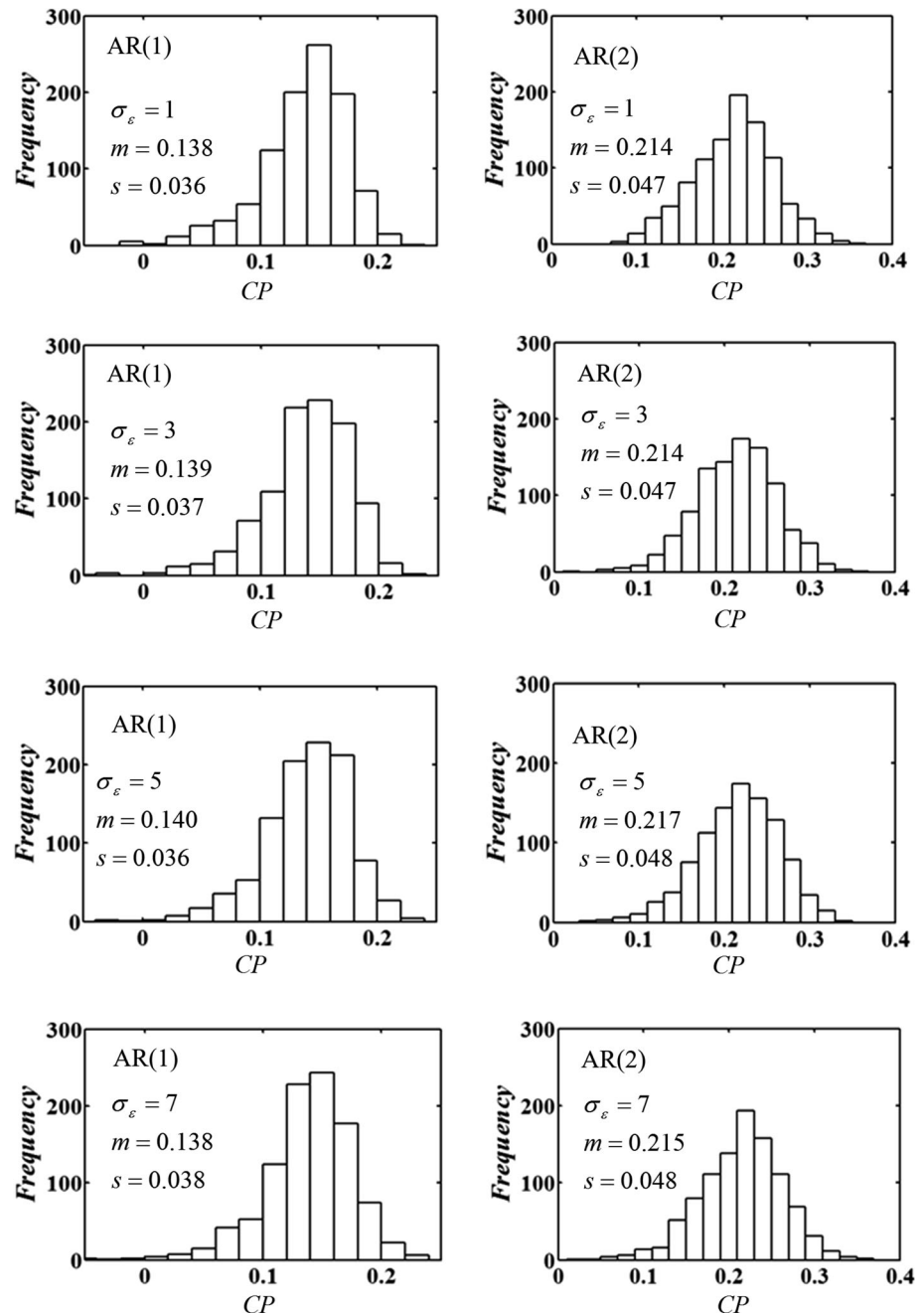
$$(18) \quad = \left(1 - \frac{SSE}{SSE_N}\right) + (2\rho_k - 1) \frac{SSE}{SSE_N} \tag{20}$$

$$= CP + (2\rho_k - 1)(1 - CP)$$

$$= 2(1 - \rho_k)CP + 2\rho_k - 1$$

Equation (20) represents the asymptotic relationship between CE and CP of any k -step lead time forecasting

Fig. 8 Histograms of the coefficient of persistence (CP) for AR(1) and AR(2) modeling with respect to various noise variance σ_ε^2

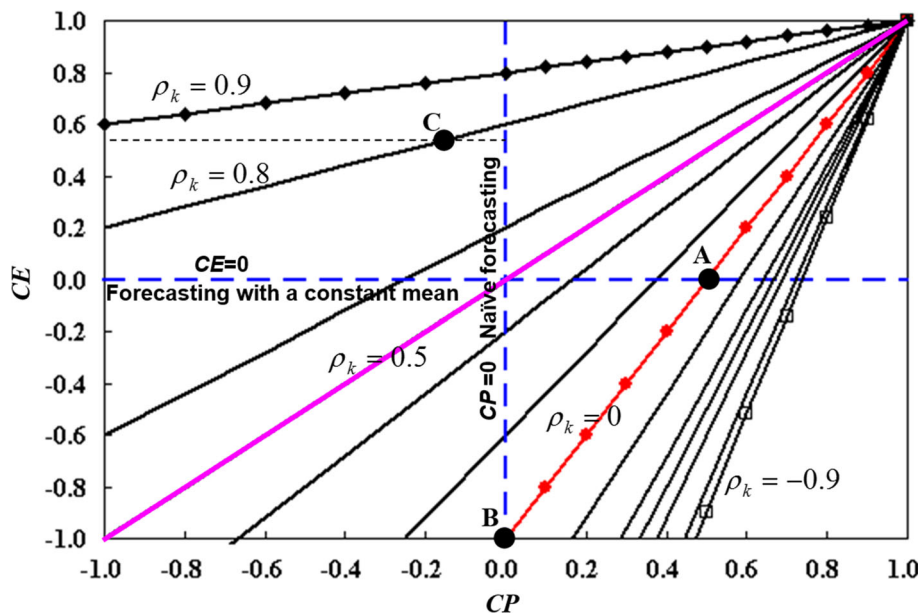


model, given a data series with a lag- k autocorrelation coefficient ρ_k . The above asymptotic relationship is illustrated in Fig. 9 for various values of lag- k autocorrelation coefficient ρ_k .

Given a data series with a specific lag- k autocorrelation coefficient, various models can be adopted for k -step lead time forecasting. Equation (20) indicates that, although the performances of these forecasting models may differ significantly, their corresponding (CE, CP) pairs will all fall on or near a specific line determined by ρ_k of the data series, as long as the data series is long enough. For example, given a data series with $\rho_1 = 0$, one-step lead

time forecasting with the constant mean ($CE = 0$) results in $CP = 0.5$ (point A in Fig. 9). Alternatively, if one chooses to conduct naïve forecasting ($CP = 0$) for the same data series, it yields $CE = -1.0$ (point B in Fig. 9). For data series with $\rho_k < 0.5$, k -step lead time forecasting with a constant mean (i.e. $CE = 0$) is superior to the naïve forecasting since the former always yields positive CP values. On the contrary, for data series with $\rho_k > 0.5$, the naïve forecasting always yields positive CE values and thus performs better than forecasting with a constant mean. Hereinafter, the CE – CP relationship of Eq. 20 will be referred to as the *sample-dependent* (or *data-dependent*)

Fig. 9 Asymptotic relationship between CE and CP for data series of various lag- k autocorrelation coefficients ρ_k . ($\rho_k = 0.9, 0.8, 0.6, 0.5, 0.4, 0.2, 0, -0.2, -0.4, -0.5, -0.6, -0.8,$ and -0.9 .)



CE – CP relationship since a sample series has a unique value of ρ_k which completely determines the CE – CP relationship. It can also be observed that the slope in Eq. 20 is smaller (or larger) than 1, if ρ_k exceeds (or is lower than) 0.5. Data series with significant persistence (high ρ_k values, such as flood flow series) are associated with very gradual CE – CP slopes. The above observation explains why CP is more sensitive than CE in Figs. 7 and 8. Thus, for real-time flood forecasting or applications of similar nature, CP is a more sensitive and suitable criterion than CE .

The asymptotic CE – CP relationship can be used to determine whether a specific CE value, for example $CE = 0.55$, can be considered as having acceptable model performance. The CE -based model performance rating recommended by Moriasi et al. (2007) does not take into account the autocorrelation structure of the data series under investigation, and thus may result in misleading recommendations. This can be explained by considering a data series with significant persistence or high lag-1 autocorrelation coefficient, say $\rho_1 = 0.8$. Suppose that a forecasting model yields a CE value of 0.55 (see point C in Fig. 9). With this CE value, performance of the model is considered satisfactory according to the performance rating recommended by Moriasi et al. (2007). However, with $\rho_1 = 0.8$ and $CE = 0.55$, it corresponds to a negative CP value ($CP = -0.125$), indicating that the model performs even poorer than the naïve forecasting, and thus should not be recommended. More specifically, if one considers naïve forecasts as the benchmark series, all one-step lead time forecasting models yielding CE values lower than $2\rho_1 - 1$ are inferior to naïve forecasting and cannot be recommended. We have found in the literature that many flow forecasting applications resulted in CE values varying

between 0.65 and 0.85. With presence of high persistence in flow data series, it is likely that not all these models performed better than the naïve forecasting.

As demonstrated in Fig. 7, variation of CE values of individual events enables us to assess the uncertainties in model performance. However, there were real-time flood forecasting studies that conducted model performance evaluation with respect to artificial continuous series of several independent events. A single CE or CP value was then calculated from the multi-event artificial series. CE values based on such artificial series cannot be considered as a measure of overall model performance with respect to all events.

For models having satisfactory performance (for example, $CE > 0.5$ for individual events), $\sum_{t=1}^n (Q_t - \hat{Q}_t)^2$ (the numerator in Eq. 9) is much smaller than $\sum_{t=1}^n (Q_t - \bar{Q})^2$ (the denominator) for all individual events. Thus, if CE is calculated for the multi-event artificial series, increase in the numerator of Eq. 9 will generally be smaller than increase in the denominator, making the resultant CE to be higher than most event-based CE values. Thus, using the CE or CP value calculated from a long artificial multi-event series may lead to inappropriate conclusions of model performance evaluation. We shall show examples of such misinterpretation in the Sect. 8.

6 Developing a CE – CP coupled MPE criterion

Another essential concern of model performance evaluation for flow forecasting is the choice of benchmark series. The benchmark should be simple, such that every hydrologist can

understand its explanatory power and, therefore, appreciate how much better the actual hydrological model is (Moussa 2010). The constant mean series and the naïve-forecast series are the benchmark series for the CE and CP criteria, respectively. Although model performance evaluations with respect to both series are easily understood and can be conveniently implemented, they only provide minimal advantages when applied to high persistence data series such as flow or stage data. Schaeffli and Gupta (2007) also argue that definition of an appropriate baseline for model performance, and in particular, for measures such as the CE values, should become part of the ‘best practices’ in hydrologic modelling. Considering the high persistence nature in flow data series, we suggest the autoregressive model $AR(p)$ be considered as the benchmark for performance evaluation of other flow forecasting models. From our previous experience in flood flow analysis and forecasting, we propose using $AR(2)$ model for benchmark comparison.

The bench coefficient G_{bench} suggested by Seibert (2001) provides a clear indication about whether the benchmark model performs better than the model under consideration. G_{bench} is negative if the model performance is poor than the benchmark, zero if the model performs as well as the benchmark, and positive if the model is superior, with a highest value of one for a perfect fit. In order to advocate using more rigorous benchmarks for model performance evaluation, we developed a CE – CP coupled MPE criterion with respect to the $AR(1)$ and $AR(2)$ models for one-step lead time forecasting. Details of the proposed CE – CP coupled criterion are described as follows.

The *sample-dependent* CE – CP relationship indicates that different forecasting models can be applied to a given data series (with a specific value of ρ_1 , say ρ^*), and the resultant (CE, CP) pairs will all fall on a line defined by Eq. 20 with $\rho_1 = \rho^*$. In other words, points on the asymptotic line determined by $\rho_1 = \rho^*$ represent performances of different forecasting models which have been applied to the given data series. Using the $AR(1)$ or $AR(2)$ model as the benchmark for model performance evaluation, we need to identify the point on the asymptotic line which corresponds to the $AR(1)$ or $AR(2)$ model. This can be achieved by the following derivations.

An $AR(1)$ random process is generally expressed as

$$X_t = \phi_1 X_{t-1} + \varepsilon_t, \varepsilon_t \sim iid \quad N(0, \sigma_\varepsilon^2), |\phi_1| < 1. \quad (21)$$

with $\rho_1 = \phi_1$ and $\sigma_\varepsilon^2 = (1 - \phi_1^2)\sigma_X^2$. Suppose that the data series under investigation is originated from an $AR(1)$ random process and an $AR(1)$ model with no parameter estimation error is adopted for one-step lead time forecasting. As the length of the sample series approaches infinity, it yields

$$CE = \phi_1^2, \quad (22)$$

and

$$CP = 1 - \frac{\sigma_\varepsilon^2}{2(1 - \rho_1)\sigma_X^2} = 1 - \frac{1 - \phi_1^2}{2(1 - \phi_1)} = \frac{1 + \phi_1}{2}. \quad (23)$$

Thus,

$$CE = (1 - 2CP)^2 = 4CP^2 - 4CP + 1. \quad (24)$$

Suppose that the data series under investigation is originated from an $AR(2)$ random process and an $AR(2)$ model with no parameter estimation error is adopted for one-step lead time forecasting. It yields

$$CP = 1 - \frac{\sigma_\varepsilon^2}{2(1 - \rho_1)\sigma_X^2} = 1 - \frac{(1 + \phi_2)(1 - \phi_2 + \phi_1)}{2}, \quad (25)$$

$$\frac{2(1 - CP)}{1 + \phi_2} + \phi_2 - 1 = \phi_1. \quad (26)$$

From Eqs. 14 and 20, it yields

$$CE = \left(\frac{4}{1 - \phi_2^2} \right) CP^2 + \left(4 - \frac{8}{1 - \phi_2^2} \right) CP + \left(\frac{4}{1 - \phi_2^2} - 3 \right). \quad (27)$$

Equations (24) and (27) respectively characterize the parabolic CE – CP relationships of the $AR(1)$ and $AR(2)$ models, and are referred to as the *model-dependent* CE – CP relationships (see Fig. 10). Unlike the *sample-dependent* CE – CP relationship of Eq. 20, Eqs. 24 and 27 describe the dependence of (CE, CP) on *model parameters* (ϕ_1, ϕ_2) . The model-dependent CE – CP relationships are derived based on the assumption that the data series are truly originated from the $AR(1)$ or $AR(2)$ model, and forecastings are conducted using perfect models (correct model types and parameters). For a specific model family, say $AR(2)$, any pair of model parameters (ϕ_1, ϕ_2) defines a unique pair of (CE, CP) on a parabolic curve determined by ϕ_2 . However, in practical applications the model and parameter uncertainties are inevitable, and the resultant (CE, CP) pairs are unlikely to coincide with their theoretical points. For model performance evaluation using the 1000 simulated series of the $AR(2)$ model with $\phi_1 = 0.5$ and $\phi_2 = 0.3$ (see details in the Sect. 4), scattering of the (CE, CP) pairs based on the $AR(1)$ and $AR(2)$ forecasting models are depicted by the two ellipses in Fig. 10. The $AR(2)$ forecasting model which does not involve the model uncertainty clearly outperforms the $AR(1)$ forecasting model.

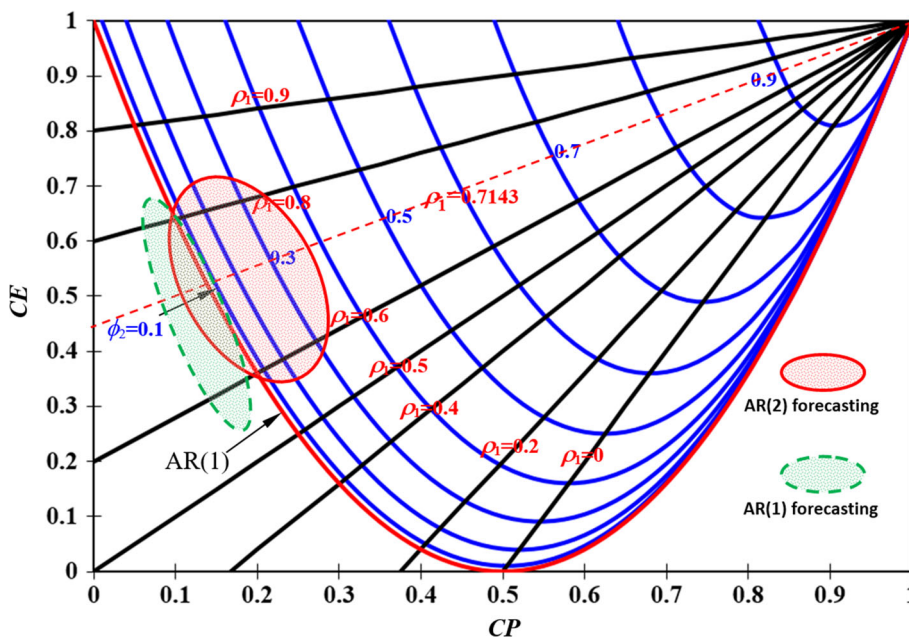


Fig. 10 Parabolic CE – CP relationships of the AR(1) and AR(2) models. The two ellipses illustrate scattering of (CE, CP) pairs for AR(1) and AR(2) forecasting of 1000 sample series of the AR(2) model $X_t = 0.5X_{t-1} + 0.3X_{t-2} + \varepsilon_t$. [See details in the Sect. 6.]

7 Bootstrap resampling for MPE uncertainties assessment

7.1 Model-based bootstrap resampling

In the previous section we used simulated AR(2) sample series to evaluate uncertainties of CE and CP . But in reality, the true properties of the sample series are never known and thus we propose to use the model-based bootstrap resampling technique to generate a large set of resampled series, and then use these resampled series for MPE uncertainties assessment. Hromadka (1997) conducted a stochastic evaluation of rainfall–runoff prediction performance based on similar concept. Details of the model-based bootstrap resampling technique (Alexeev and Tapon 2011; Selle and Hannah 2010) are described as follows.

Assuming that a sample data series $\{x_1, x_2, \dots, x_n\}$ is available, we firstly subtract the mean value (\bar{x}_n) from the sample series to yield a zero-mean series, i.e.,

$$x_t^* = x_t - \bar{x}_n, \quad t = 1, 2, \dots, n. \tag{28}$$

A set of resampled series is then generated through the following procedures:

- (1) Select an appropriate model for the zero-mean data series $\{x_t^*, t = 1, 2, \dots, n\}$ and then estimate the model parameters. In this study the AR(2) model is adopted since we focus on real-time forecasting of flood flow time series which exhibits significant

persistence. Let $\hat{\phi}_1$ and $\hat{\phi}_2$ be estimates of the AR(2) parameters, the residuals can then be calculated as

$$e_t = x_t^* - (\hat{\phi}_1 x_{t-1}^* + \hat{\phi}_2 x_{t-2}^*), \quad t = 1, \dots, n. \tag{29}$$

- (2) The residuals are then centered with respect to the residual mean (\bar{e}_n) , i.e.

$$\tilde{e}_t = e_t - \bar{e}_n, \quad t = 1, \dots, n. \tag{30}$$

- (3) A set of bootstrap residuals $(\varepsilon_t, t = 1, \dots, n)$ is obtained by re-sampling with replacement from the centered residuals $(\tilde{e}_t, t = 1, \dots, n)$.

- (4) A bootstrap resampled series $\{y_1, y_2, \dots, y_n\}$ is then obtained as

$$y_t = (\hat{\phi}_1 x_{t-1}^* + \hat{\phi}_2 x_{t-2}^* + \varepsilon_t) + \bar{x}_n, \quad t = 1, \dots, n. \tag{31}$$

7.2 Flood forecasting model performance evaluation

Hourly flood flow time series (see Fig. 11) of nine storm events observed at the outlet of the Chih-Lan River watershed in southern Taiwan were used to demonstrate the uncertainties in flood forecasting model performance based on bootstrap resampled flood flow series. The Chih-Lan River watershed encompasses an area of 110 km². All flow series have very high lag-1 autocorrelation coefficients ($\rho_1 > 0.8$) due to significant flow persistence. For each of the nine observed flow series, a total of 1000

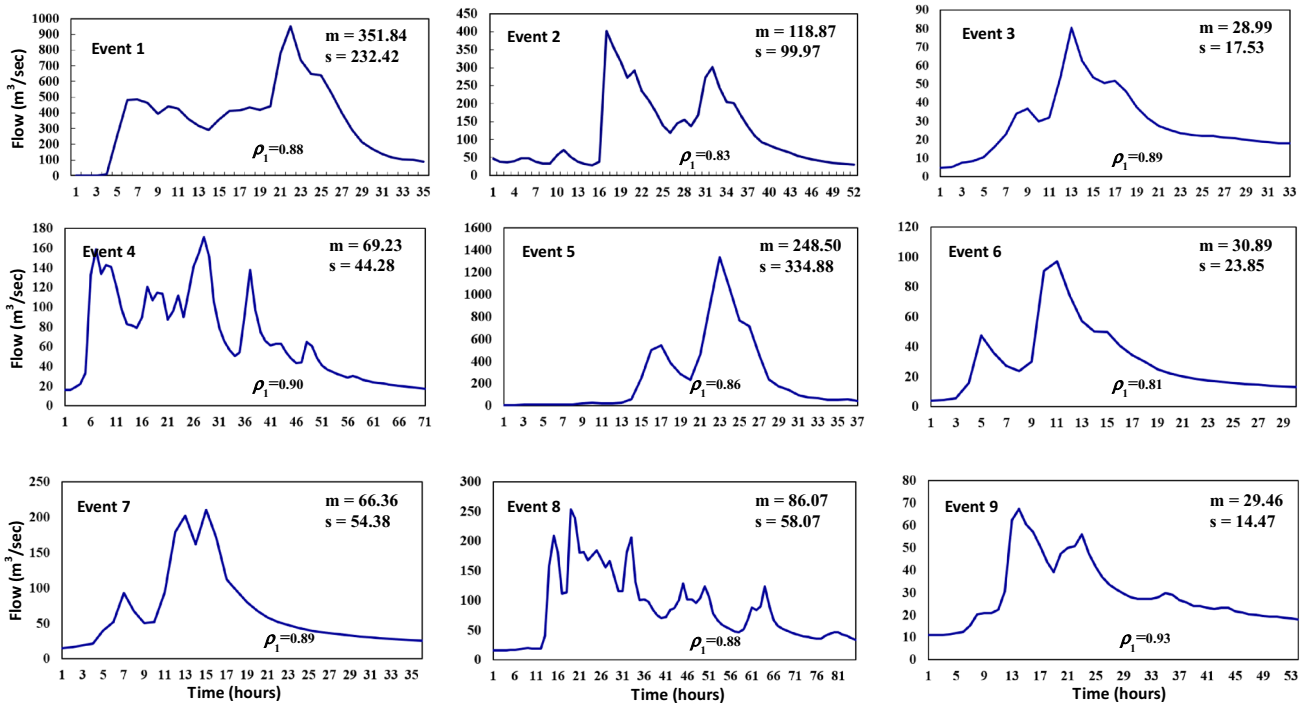


Fig. 11 Flow hydrographs of the flood events used in this study. The mean (m), standard deviation (s) and lag-1 autocorrelation coefficient (ρ_1) of individual flow series are also shown

bootstrap resampled series was generated through the model-based bootstrap resampling. These resampled series were then used for assessing uncertainties in model performance evaluation.

The artificial neural network (ANN) has been widely applied for different hydrological predictions, including real-time flood forecasting. Thus, we evaluate the model performance uncertainties of an exemplar ANN model for real-time flood forecasting, using the AR(2) model as the benchmark. In particular, we aim to assess the capability of the exemplar ANN model for real-time forecasting of random processes with high persistence, such as flood flow series. In our flood forecasting model performance evaluation, we only consider flood forecasting of one-step (1 h) lead time. For small watersheds, the times of concentration usually are less than a few hours, and thus flood forecasts of lead time longer than the time of concentration are less useful. Besides, if the performance of the one-step lead time forecasts is not satisfactory, forecasts of longer lead time (multiple-step lead time) will not be necessary.

For forecasting with an AR(2) model, the nine observed flood flow series were divided into two datasets. The calibration dataset is comprised of 6 events (events 1, 2, 3, 4, 7 and 9) and the test dataset consists of the remaining three events. Using flow series in the calibration dataset, flood flows at the watershed outlet can be expressed as the following AR(2) random process:

$$x_t = 7.3171 + 1.2415x_{t-1} - 0.3173x_{t-2} + \varepsilon_t, \varepsilon_t \sim iid \\ N(0, \sigma_\varepsilon = 43.96 \text{ m}^3/\text{s}) \quad (32)$$

Thus, the one-step lead time flood forecasting model for the watershed was established as

$$\hat{x}_t = 7.3171 + 1.2415x_{t-1} - 0.3173x_{t-2} \quad (33)$$

The above equation was then applied to the 1000 bootstrap resampled series of each individual event for real-time flood forecasting. Figure 12 shows scattering of (CE , CP) of the resampled series of individual events. The means and standard deviations of CE and CP are listed in Table 2.

For ANN flood flow forecasting, an exemplar back-propagation network (BPN) model with one hidden layer of two nodes was adopted in this study. The BPN model uses three observed flows (x_t , x_{t-1} , x_{t-2}) in the input layer for flood forecasting of x_{t+1} . An ANN model needs to be trained and validated. Thus, the calibration dataset of the AR(2) modeling was further divided into two groups. Events 1, 4 and 9 were used for training and events 2, 3 and 7 were used for validation. After completion of training and validation, the BPN model structure and weights of the trained model were fixed and applied to the bootstrap resampled series of individual events. Figure 13 shows scattering of (CE , CP) based on BPN forecasts of

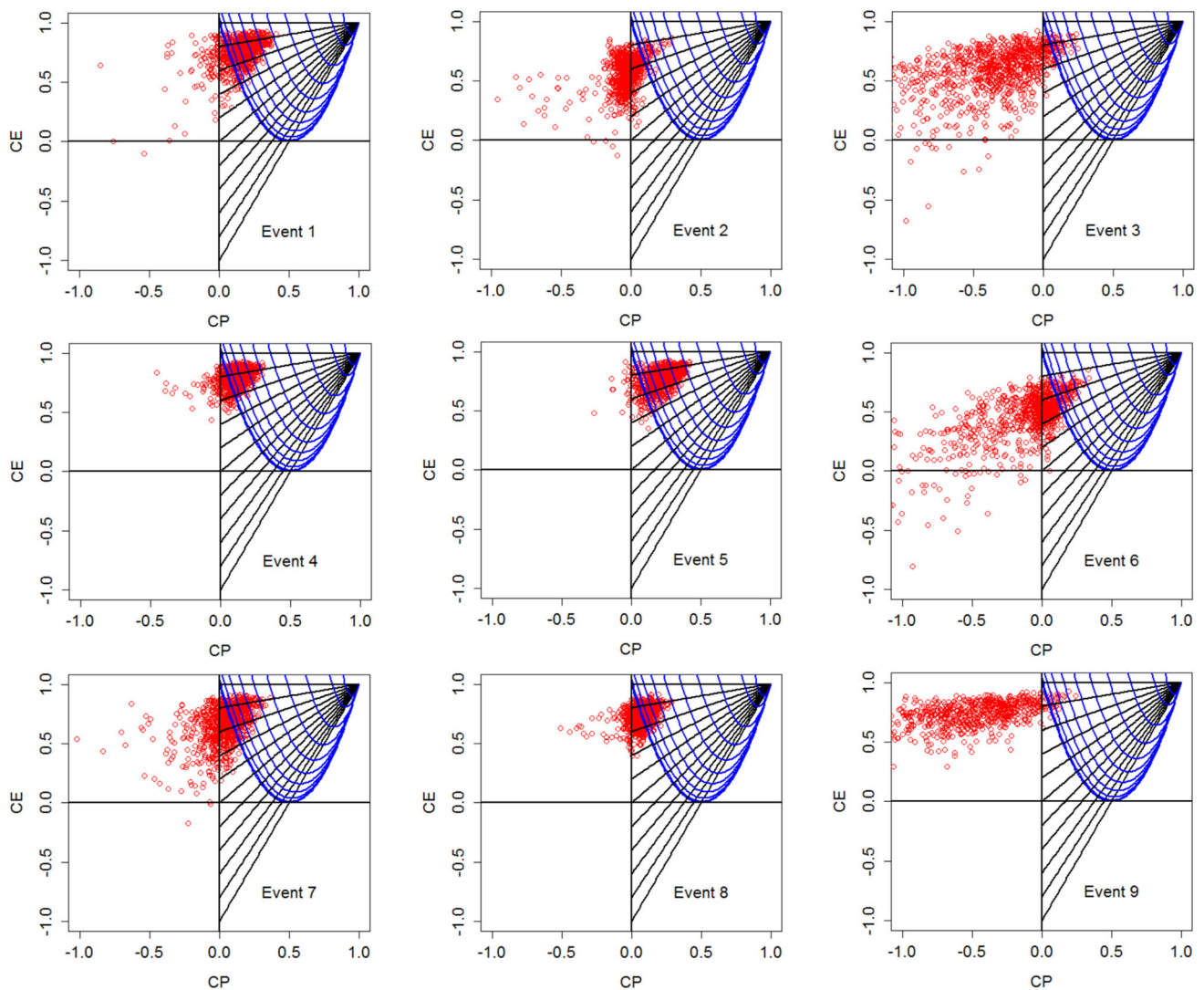


Fig. 12 Model performance uncertainties in terms of (CE , CP). The linear and parabolic CE – CP relationships have been illustrated in Figs. 9 and 10. [AR(2) model for real-time flood forecasting.]

resampled series. The means and standard deviations of CE and CP of BPN forecasts are also listed in Table 3.

With the very simple and pre-calibrated AR(2) model, CE values of most resampled-series are higher than 0.5 and can be considered in the ratings of good to very good according to Moriasi et al. (2007). Whereas a significant portion of the bootstrap resampled series of events 2, 3, 6 and 9 are associated with negative CP values, suggesting that AR(2) forecasting for these events are inferior to the naïve forecasting. Although the AR(2) and BPN models yielded similar (CE , CP) scattering patterns for resampled series of all individual events, the BPN forecasting model yielded negative average CP values for six events, comparing to four events for the AR(2) model.

Resampled-series-wise comparison of (CE , CP) of the two models was also conducted. For each resampled series, CE and CP values of the AR(2) and BPN models were

compared. The model with higher values is considered superior to the other, and the percentages of model superiority for AR(2) and BPN were calculated and shown in Table 4. Among the nine events, AR(2) model achieves dominant superiority for four events (events 2, 4, 7 and 8), whereas the BPN model achieves dominant superiority for events 3 and 9 only. Overall, the AR(2) model is superior to the BPN model for 61.5 and 54.4 % of all resampled series in terms of CE and CP , respectively. It is also worthy to note that the AR(2) model is superior in terms of CE and CP simultaneously for nearly half (48.7 %) of all resampled series. Han et al. (2007) assessed the uncertainties in real-time flood forecasting with ANN models and found that ANN models are uncompetitive against a linear transfer function model in short-range (short lead time) predictions and should not be used in operational flood forecasting owing to their complicated calibration process.

Table 2 Mean and standard deviation of *CE* and *CP* of the resampled series of individual events [AR(2) forecasting]

Event	<i>CE</i>		<i>CP</i>		Remark
	Mean	SD	Mean	SD	
1	0.7549	0.1190	0.1542	0.1227	Calibration
2	0.5600	0.1362	-0.0536	0.3041	Calibration
3	0.5369	0.3292	-0.5377	0.8172	Calibration
4	0.7858	0.0743	0.1121	0.0824	Calibration
5	0.7773	0.0909	0.2215	0.0952	Test
6	0.4311	0.2802	-0.2326	0.5939	Test
7	0.6666	0.1581	0.0372	0.1498	Calibration
8	0.7354	0.0824	0.0680	0.0780	Test
9	0.6050	0.2892	-1.292	1.3896	Calibration

The results of our evaluation are consistent with such findings and reconfirm the importance of taking into account the persistence in flood series in model performance evaluation.

Considering the magnitude of flows (see Fig. 11), the BPN model seems to be more superior for events of lower flows (events 3 and 9) whereas the AR(2) model has dominant superiority for events of median flows (events 2, 4, 7 and 8). For events of higher flows (events 1 and 5), performance of the two models are similar. Figure 14 demonstrates that the average *CE* and *CP* values tend to increase with mean flows of individual flood events. The dependence is apparently more significant between the average *CP* and mean flow of the event. This result is consistent with previous findings that *CP* is more sensitive

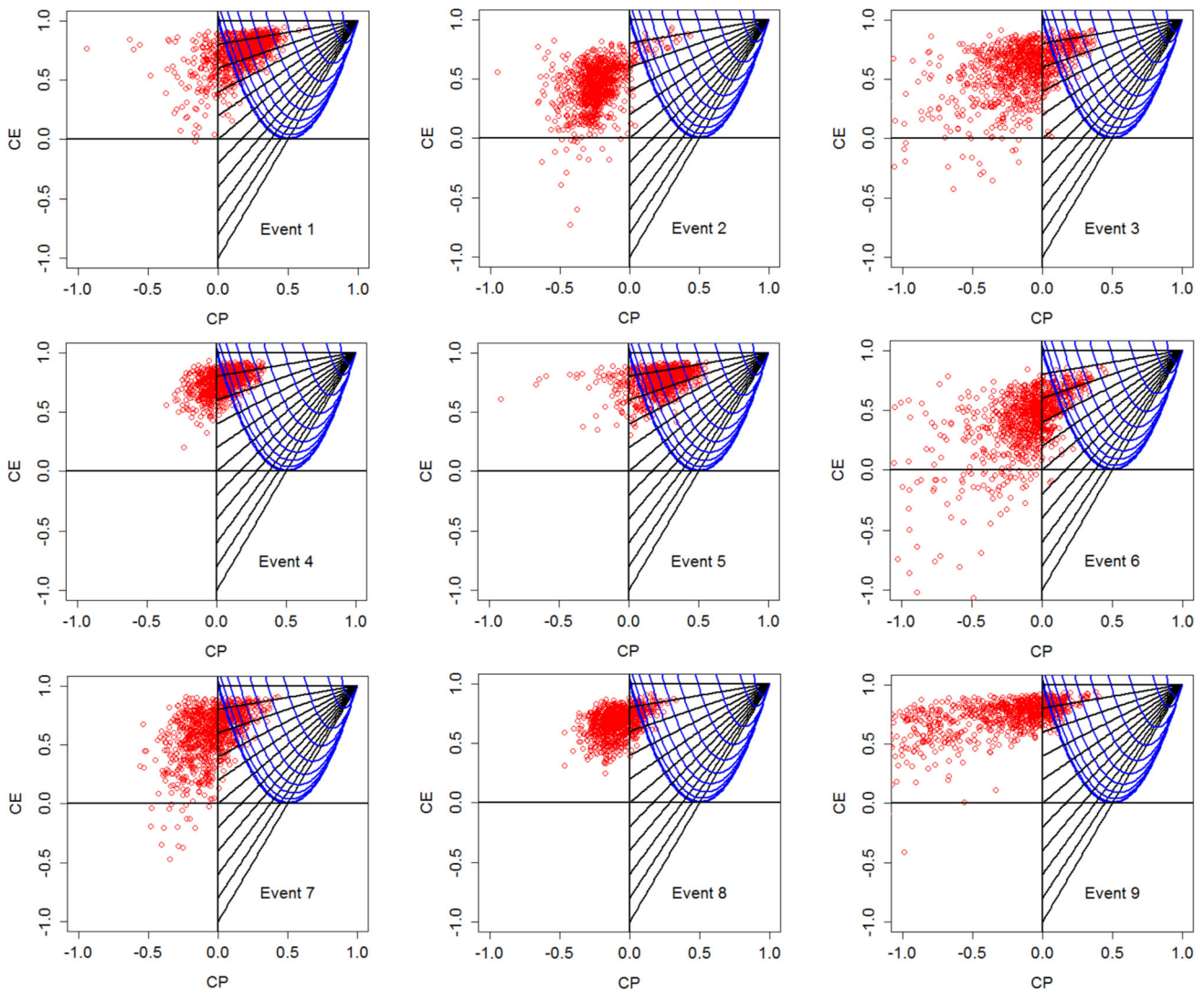


Fig. 13 Model performance uncertainties in terms of (*CE*, *CP*). The linear and parabolic *CE*–*CP* relationships have been illustrated in Figs. 9 and 10. [BPN model for real-time flood forecasting.]

Table 3 Mean and standard deviation of *CE* and *CP* of the resampled series of individual events [BPN forecasting]

Event	<i>CE</i>		<i>CP</i>		Remark
	Mean	SD	Mean	SD	
1	0.7320	0.1441	0.1651	0.1731	Training
2	0.4471	0.1901	−0.2330	0.1569	Validation
3	0.5742	0.2804	−0.1944	0.4578	Validation
4	0.7577	0.0942	0.0493	0.1139	Training
5	0.7774	0.0965	0.2301	0.1825	Test
6	0.4274	0.2599	−0.1144	0.2891	Test
7	0.6043	0.2121	−0.0430	0.1631	Validation
8	0.6796	0.1054	−0.0804	0.1161	Test
9	0.7111	0.1882	−0.4204	0.6270	Training

Table 4 Sample-wise (*CE*, *CP*) comparison

Event	Ratio of AR(2) superiority ^a			Ratio of BPN superiority ^a		
	<i>CE</i>	<i>CP</i>	<i>CE&CP</i>	<i>CE</i>	<i>CP</i>	<i>CE&CP</i>
1	0.610	0.408	0.336	0.390	0.592	0.318
2	0.916	0.942	0.901	0.084	0.058	0.043
3	0.349	0.094	0.052	0.651	0.906	0.609
4	0.839	0.815	0.744	0.161	0.185	0.090
5	0.455	0.353	0.290	0.545	0.647	0.482
6	0.576	0.490	0.404	0.424	0.510	0.338
7	0.766	0.805	0.711	0.234	0.195	0.140
8	0.951	0.969	0.941	0.049	0.031	0.021
9	0.076	0.023	0.001	0.924	0.977	0.902
Overall	0.615	0.544	0.487	0.385	0.456	0.327

^a The ratio of model superiority represents the proportion of the resampled series that a model (AR(2) or BPN) achieves higher *CE* or *CP* values than the other

than *CE*, and is a more suitable criterion for real-time flood forecasting.

It is also worthy to note that a few studies had evaluated the performance of forecasting models using *CE* calculated from multi-event artificial series (Chang et al. 2004; Chiang et al. 2007; Chang et al. 2009; Chen et al. 2013; Wei, 2014). To demonstrate the effect of using *CE* calculated from multi-event artificial series for performance evaluation of *event-based* forecasting (such as flood forecasting) models, *CE* and *CP* values calculated with respect to individual flood events and multi-event artificial series are shown in Fig. 15. The artificial flow series combines observed (or AR(2)-forecast) flow hydrographs of Event-1 to Event-5 in Fig. 11. *CE* value of the multi-event artificial series is higher than *CE* values of *any* individual events. Particularly, in contrast to the high *CE* value

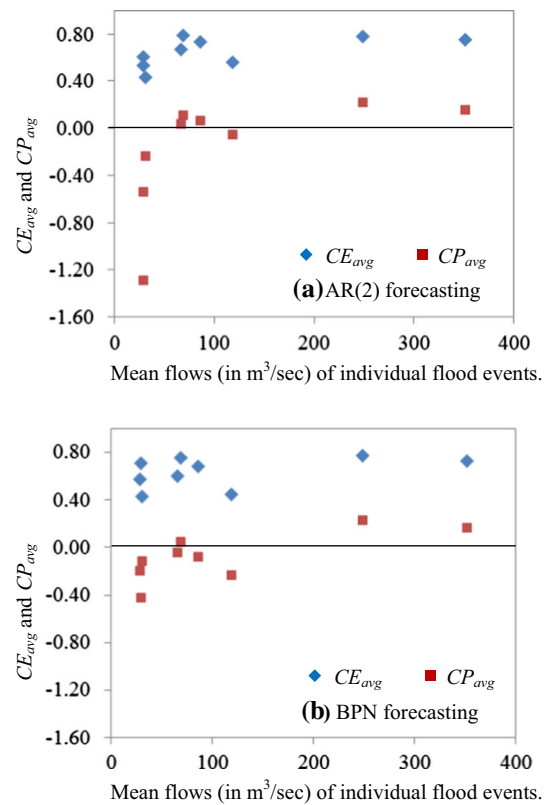


Fig. 14 Model performance evaluation (in terms of CE_{avg} and CP_{avg}) with respect to mean flows of individual flood events. **a** AR(2) forecasting. **b** BPN forecasting. [Note: CE_{avg} and CP_{avg} are average values of *CE* and *CP* of the 1000 bootstrap resampled series.]

(0.879) of the artificial series, Event-2 and Event-3 have lower *CE* values (0.665 and 0.668, respectively). Although the artificial series yields a positive *CP* value (0.223), Event-2 and Event-3 are associated with negative *CP* values (−0.009 and −0.176, respectively). We have also found that long artificial series consisting of more individual flood events are very likely to result in very high *CE* values (for examples, between 0.93 and 0.98, Chen et al., 2013) for short lead-time forecast. We argue that for such studies *CE* values of individual flood events could be lower and some events were even associated with negative event-specific *CP* values.

Results in Fig. 15 show that *CE* value of the multi-event series is higher than all event-based *CE* values. However, under certain situations, for example forecasts of higher flows are less accurate, *CE* value of the multi-event series can be smaller than only a few event-based *CE* values. To demonstrate such a situation, we manually adjusted the AR(2) forecasts for two events (event 1 and event 5) with higher flood flows such that their forecasts are less accurate than those of the other three events. We then recalculated *CE* values for individual events and the multi-event series, and the results are shown in Fig. 16. With less accurate

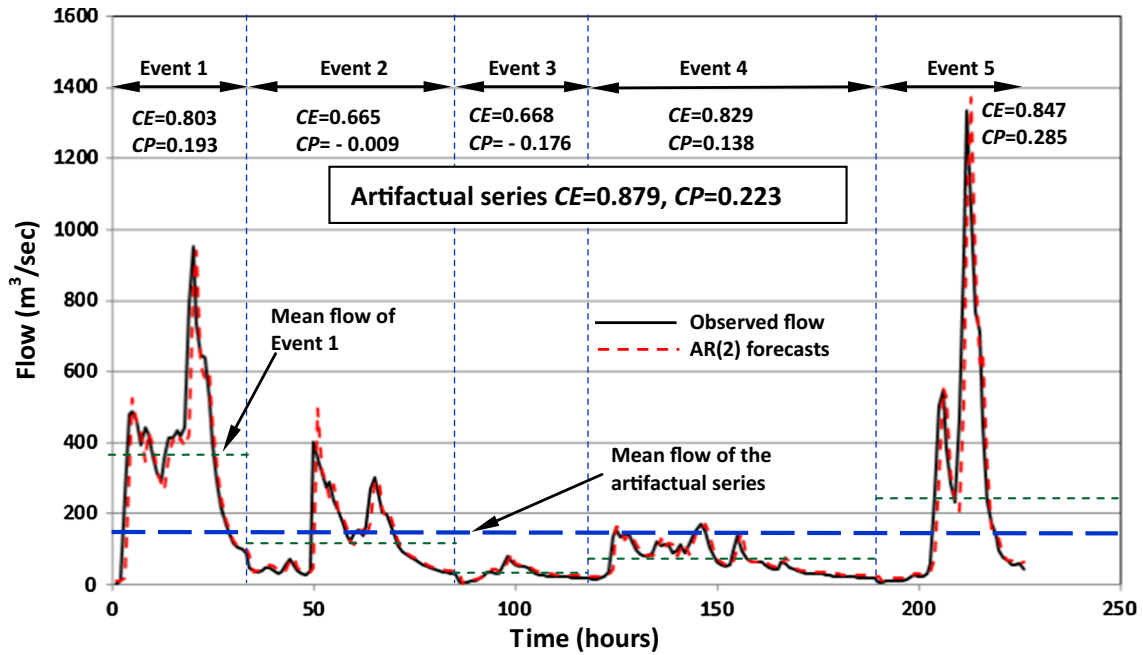


Fig. 15 Comparison of (*CE*, *CP*) values with respect to individual events and (*CE*, *CP*) of the multi-event artificial series. Forecasts are based on an AR(2) model. The artificial series yielded higher *CE*

value than any individual event. *CP* of the artificial series is positive whereas two events are associated with negative *CP* values

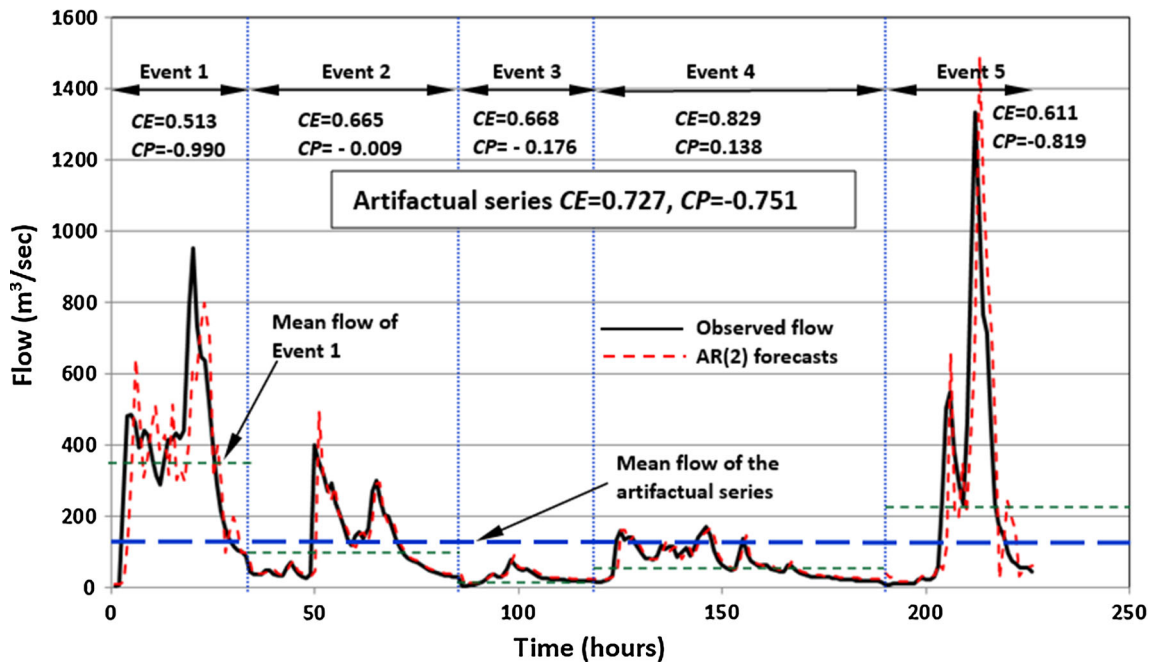


Fig. 16 Comparison of (*CE*, *CP*) values with respect to individual events and (*CE*, *CP*) of the multi-event artificial series. Forecasts of events 2, 3, and 4 are based on an AR(2) model. Forecasts of event 1 and 5 were manually adjusted from AR(2) forecasts to become less

accurate. The multi-event artificial series yielded higher *CE* value than all individual event, except event 4. *CP* values were negative for the artificial series and four individual events

forecasts for events 1 and 5, *CE* values of the two events and the multi-event artificial series were reduced. *CE* value of the multi-event artificial series (0.727) became

smaller than *CE* of event 4 (0.829). However, the multi-event *CE* value was still larger than event-based *CE* values for 4 of the 5 events. It can also be observed that the multi-

event CP value changed from 0.223 to -0.751 . This demonstrates that CP is a more powerful test of model performance (i.e. capable of clearly indicating poor model performance) than CE . In this example, forecasts of events 1 and 5 (having higher flows) were manually adjusted to make them less accurate. However, for models which yield similar forecast performance for low to high flood events (i.e. having consistent model performance), we believe that CE value of the artificial multi-event series is likely to be higher than all event-based CE values.

We have also found a few studies that aimed to simulate or continuously forecast daily or monthly flow series over a long period. Most of such applications are related to water resources management or for the purpose of understanding the long-term hydrological behaviors such as snow-melt runoff process and baseflow process (Schreider et al. 1997; Dibike and Coulibaly 2007; Chiew et al. 2014; Wang et al. 2014; Yen et al. 2015). For such applications, long-term simulation or forecasts of flow series were required and CE and CP measures were calculated for flow series spanning over one-year or multiple-year periods. However, in contrast to these aforementioned studies, the work of real-time flood forecasting is *event-based* and the model performance can vary from one event to another, it is therefore imperative for researchers and practitioners to look into the model performance uncertainties. A single CE or CP value derived from a multi-event artificial series does not provide a multi-event *overall* evaluation and may actually disguise the real capability of the proposed model. Thus, CE or CP value derived from a multi-event artificial series should not be used for event-based forecasting practices.

8 MPE for multiple-step lead time flood forecasting

In the previous section, we only consider one-step lead time forecasting models. There are also studies (for example, Chen et al. 2013) that aimed to develop multiple-step lead time flood forecasting models. Using CP as the MPE criterion for multiple-step lead time flood forecasting deserves a careful look.

For a k -step lead time flood forecasting, the sample-dependent asymptotic CE – CP relationship is determined by ρ_k of the data series. Generally speaking, the flow persistence and ρ_k decrease as the time lag k increases. For large enough lead time steps (for examples, 4-step or 6-step lead time forecasts), ρ_k becomes lower and the naïve forecasting models can be expected to yield poor performance. Thus, it is possible to yield positive CP values for multiple-step lead time forecasts, whereas CP value of one-step lead time forecasts of the same model is negative. For

such cases, it does not imply that the model performs better in multiple-step lead time than in one-step lead time. Instead, it's the naïve forecasting model which performs much worse in multiple-step lead time. Since ρ_k of flood flow series often reduces to lower than 0.6 for $k \geq 3$, we recommend model performance evaluation using CP be limited to one or two-step lead time flood forecasting. Using CP for performance evaluation of multiple-step forecasting should be exercised with extra caution. Especially we warn of using CP values derived from *multi-event artificial series* for model performance evaluation of *multiple-step lead time* flood forecasting. Such practices may further exacerbate the misleading conclusions about the real forecasting capabilities of the proposed models.

9 Summary and conclusions

We derived the sample-dependent and AR model-dependent asymptotic relationships between CE and CP . Considering the temporal persistence in flood flow series, we suggest using AR(2) model as the benchmark for event-based flood forecasting model performance evaluation. Given a set of flow hydrographs (test events), a CE – CP coupled model performance evaluation criterion for *event-based* flood forecasting is proposed as follows:

- (1) Calculate CE and CP of the proposed model and the AR(2) model for one-step lead time flood forecasting. A model yielding negative CP values is inferior to the naïve forecasting and cannot be considered for real-time flood forecasting.
- (2) Compare CP values of the proposed model and the AR(2) model. If CP of the proposed model is lower than CP of the AR(2) model, the proposed model is inferior to the AR(2) model.
- (3) If the proposed model yields positive and higher-than-AR(2) CP values, evaluate its CE values. Considering the significant lag-1 autocorrelation coefficient ($\rho_1 > 0.8$) for most of the flood flow series and forecasting capability of the AR(2) model, we suggest that the CE value should exceed 0.70 in order for the proposed model to be acceptable for real-time flood forecasting. However, for flood forecasting of larger watersheds, flow series at the watershed outlet may have even higher lag-1 autocorrelation coefficients and the threshold CE value should be raised accordingly (for example, $CE > 0.85$ for $\rho_1 > 0.9$).
- (4) The above steps provide a first phase event-based model performance evaluation. It is also advisable to conduct bootstrap resampling of the observed flow series and calculate the bootstrap-series average (CE ,

CP) values of the proposed model and the AR(2) model for individual flood events. The bootstrap-series average (CE , CP) values can then be used to evaluate the model performance using the same criteria in steps 1–3.

- (5) Multiple-step lead time flood forecasting should be considered only if the proposed model yields acceptable performance of one-step lead time forecasting through the above evaluation.

In addition to the above CE – CP coupled MPE criterion for real-time flood forecasting, a few concluding remarks are also given as follows:

- (1) Both CE and CP are goodness-of-fit measures of the model forecasts to the observed flow series. With significant flow persistence, even the naïve forecasting can achieve high CE values for real-time flood forecasting. Thus, CP should be used to screen out models which yield serious lagged-forecast results.
- (2) For any given data series, there exists an asymptotic linear relationship between CE and CP of the model forecasts. For k -step lead time forecasting, the relationship is dependent on the lag- k autocorrelation coefficient.
- (3) For AR(1) and AR(2) data series, the model-dependent asymptotic relationships of CE and CP can be represented by parabolic curves which are dependent on AR parameters.
- (4) Flood flow series generally have lag-1 autocorrelation coefficient higher than 0.8 and thus the AR model can easily achieve reasonable performance of real-time flood forecasting. Comparing to forecasting with a constant mean and naïve forecasting, the simple and well-known AR(2) model is a better choice of benchmark reference model for real-time flood forecasting. Flood forecasting models are recommended only if their performances (based on the above CE – CP coupled criterion) are superior to the AR(2) model.
- (5) A single CE or CP value derived from a multi-event artificial series by no means provides a multi-event *overall* evaluation and may actually disguise the real capability of the proposed model. Thus, CE or CP value derived from a multi-event artificial series should never be used in any event-based forecasting practices.
- (6) It is possible for a model to yield positive CP values for multiple-step lead time forecasts, whereas CP value of one-step lead time forecasts of the same model is negative. For such cases, it does not imply that the model performs better in multiple-step lead time than in one-step lead time.

In concluding this paper, we like to cite the following comment of Seibert (2001) which not only is truthful but thought-provoking:

Obviously there is the risk of discouraging results when a model does not outperform some simpler way to obtain a runoff series. But if we truly wish to assess the worth of models, we must take such risks. Ignorance is no defense.

Acknowledgments We gratefully acknowledge the funding support of the Ministry of Science and Technology of Taiwan for a research project (NSC-99-2911-I-002-125) which led to results presented in this paper. We declare that there is no conflict of interest with respect to any analysis, result presentation and conclusions in the paper. We also thank three anonymous reviewers for their constructive and insightful comments which led to a much improved presentation of this paper.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Alexeev V, Tapon F (2011) Testing weak form efficiency on the Toronto Stock Exchange. *J Empir Financ* 18:661–691
- Ancil F, Rat A (2005) Evaluation of neural network streamflow forecasting on 47 watersheds. *J Hydrol Eng* 10:85–88
- Andrews DWK, Chen HY (1994) Approximately median-unbiased estimation of autoregressive models. *J Bus Econ Stat* 12(2):187–204
- ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management Committee (1993) Criteria for evaluation of watershed models. *J Irrig Drain Eng* 119(3):429–442
- ASCE Task Committee on Application of the Artificial Neural Networks in Hydrology (2000) Application of the artificial neural networks in hydrology I: preliminary concepts. *J Hydrol Eng* 5(2):115–123
- Bergström S (1976) Development and application of a conceptual runoff model for Scandinavian catchments. Report RHO 7, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden
- Bergström S, Forsman A (1973) Development of a conceptual deterministic rainfall–runoff model. *Nord Hydrol* 4:147–170
- Beven KJ (1989) Changing ideas in hydrology: the case of physically based models. *J Hydrol* 105:157–172
- Beven KJ (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. *Adv Water Resour* 16:41–51. doi:10.1016/0309-1708(93)90028-E
- Beven KJ (2006) A manifesto for the equifinality thesis. *J Hydrol* 320:18–36. doi:10.1016/j.jhydrol.2005.07.007
- Beven KJ, Binley AM (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrol Process* 6:279–298. doi:10.1002/hyp.3360060305
- Calvo B, Savi F (2009) Real-time flood forecasting of the Tiber River in Rome. *Nat Hazards* 50:461–477

- Chang LC, Chang FJ, Chiang TM (2004) A two-step-ahead recurrent neural network for stream-flow forecasting. *Hydrol Process* 18:81–92
- Chang LC, Chang FJ, Wang YP (2009) Auto-configuring radial basis function networks for chaotic time series and flood forecasting. *Hydrol Process* 23:2450–2459
- Chen PA, Chang LC, Chang FJ (2013) Reinforced recurrent neural networks for multi-step-ahead flood forecasts. *J Hydrol* 497:71–79
- Chiang YM, Hsu KL, Chang FJ, Hong Y, Sorooshian S (2007) Merging multiple precipitation sources for flash flood forecasting. *J Hydrol* 340:183–196
- Chiew FHS, Potter NJ, Vaze J, Petheram C, Zhang L, Teng J, Post DA (2014) Observed hydrologic non-stationarity in far south-eastern Australia: implications for modelling and prediction. *Stoch Environ Res Risk Assess* 28:3–15
- Cloke HL, Pappenberger F (2009) Ensemble flood forecasting: a review. *J Hydrol* 375:613–626
- Corzo G, Solomatine D (2007) Baseflow separation techniques for modular artificial neural network modelling in flow forecasting. *Hydrol Sci J* 52(3):491–507
- Coulibaly P, Evora ND (2007) Comparison of neural network methods for infilling missing daily weather records. *J Hydrol* 341:27–41
- Dibike YB, Coulibaly P (2007) Validation of hydrological models for climate scenario simulation: the case of Saguenay watershed in Quebec. *Hydrol Process* 21:3123–3135
- Du J, Xie H, Hu Y, Xu Y, Xu CY (2009) Development and testing of a new storm runoff routing approach based on time variant spatially distributed travel time method. *J Hydrol* 369:44–54
- Gupta HV, Sorooshian S, Yapo PO (1999) Status of Automatic calibration for hydrologic models: comparison with multilevel expert calibration. *J Hydrol Eng* 4:135–143
- Han D, Kwong T, Li S (2007) Uncertainties in real-time flood forecasting with neural networks. *Hydrol Process* 21(2):223–228
- Harmel RD, Smith PK (2007) Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. *J Hydrol* 337:326–336
- Hromadka TV II (1997) Stochastic evaluation of rainfall–runoff prediction performance. *J Hydrol Eng* 2(4):188–196
- Kasiswathanan KS, Sudheer KP (2013) Quantification of the predictive uncertainty of artificial neural network based river flow forecast models. *Stoch Environ Res Risk Assess* 27:137–146
- Kitanidis PK, Bras RL (1980) Real-time forecasting with a conceptual hydrologic model, 2, applications and results. *Water Resour Res* 16(6):1034–1044
- Kuczera G (1997) Efficient subspace probabilistic parameter optimization for catchment models. *Water Resour Res* 33(1):177–185
- Kuczera G, Mroczkowski M (1998) Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Water Resour Res* 34(6):1481–1489
- Labat D, Ababou R, Mangin A (1999) Linear and nonlinear input/output models for karstic springflow and flood prediction at different time scales. *Stoch Environ Res Risk Assess* 13:337–364
- Lauzon N, Anctil F, Baxter CW (2006) Clustering of heterogeneous precipitation fields for the assessment and possible improvement of lumped neural network models for streamflow forecasts. *Hydrol Earth Syst Sci* 10:485–494
- Lee G, Tachikawa Y, Sayama T, Takara K (2012) Catchment responses to plausible parameters and input data under equifinality in distributed rainfall–runoff modeling. *Hydrol Process* 26:893–906. doi:10.1002/hyp.8303
- Legates DR, McCabe GJ Jr (1999) Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour Res* 35(1):233–241
- Lin GF, Wu MC, Chen GR, Tsai FY (2009) An RBF-based model with an information processor for forecasting hourly reservoir inflow during typhoons. *Hydrol Process* 23:3598–3609
- Lindström G, Johansson B, Persson M, Gardelin M, Bergström S (1997) Development and test of the distributed HBV-96 hydrological model. *J Hydrol* 201:272–288
- Markus M, Tsai CWS, Demissie M (2003) Uncertainty of weekly nitrate-nitrogen forecasts using artificial neural networks. *J Environ Eng* 129(3):267–274
- Michaud JD, Sorooshian S (1994) Comparison of simple versus complex distributed runoff models on a midsized semiarid watershed. *Water Resour Res* 30(3):593–605
- Moore RJ, Bell VA, Jones DA (2005) Forecasting for flood warning. *CR Geosci* 337:203–217
- Moriassi DN, Arnold JG, Liew MWV, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the American Society of Agricultural and Biological Engineers* 50(3):885–900
- Moussa R (2010) When monstrosity can be beautiful while normality can be ugly: assessing the performance of event-based flood models. *Hydrol Sci J* 55(6):1074–1084
- Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models. Part I. A discussion of principles. *J Hydrol* 10:282–290
- Pebesma EJ, Switzer P, Loague K (2007) Error analysis for the evaluation of model performance: rainfall–runoff event summary variables. *Hydrol Process* 21:3009–3024
- Refsgaard JC (1994) Model and data requirements for simulation of runoff and land surface processes in relation to global circulation model. In: NATO Advanced Science Institute on Global Environmental Change, Sorooshian S, Gupta HV, Rodda SC (eds) Global environmental change and land surface processes in hydrology: the trials and tribulations of modeling and measuring. Springer, Berlin, pp 169–180
- Rodriguez-Iturbe I, Valdés JB (1979) The geomorphology structure of hydrologic response. *Water Resour Res* 15(6):1409–1420
- Rodriguez-Iturbe I, González-Sanabria M, Bras RL (1982) A geomorphoclimatic theory of the instantaneous unit hydrograph. *Water Resour Res* 18(4):877–886
- Sahoo GB, Ray C, De Carlo EH (2006) Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. *J Hydrol* 327:525–538
- Sarangi A, Bhattacharya AK (2005) Comparison of artificial neural network and regression models for sediment loss prediction from Banha watershed in India. *Agric Water Manag* 78:195–208
- Sattari MT, Yurekli K, Pal M (2012) Performance evaluation of artificial neural network approaches in forecasting reservoir inflow. *Appl Math Model* 36:2649–2657
- Sauter T, Schneider C, Kilian R, Moritz M (2009) Simulation and analysis of runoff from a partly glaciated meso-scale catchment area in Patagonia using an artificial neural network. *Hydrol Process* 23:1019–1030
- Schaeffli B, Gupta HV (2007) Do Nash values have value? *Hydrol Process* 21:2075–2080. doi:10.1002/hyp.6825
- Schreider SY, Jakeman AJ, Dyer BG, Francis RI (1997) A combined deterministic and self-adaptive stochastic algorithm for streamflow forecasting with application to catchments of the Upper Murray Basin, Australia. *Environ Model Softw* 12(1):93–104
- Seibert J (1999) Conceptual runoff models—fiction or representation of reality? Acta Universitatis Uppsala, Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology
- Seibert J (2001) On the need for benchmarks in hydrological modelling. *Hydrol Process* 15:1063–1064

- Seibert J, McDonnell JJ (2002) On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration. *Water Resour Res* 38:1241. doi:[10.1029/2001WR000978](https://doi.org/10.1029/2001WR000978)
- Selle B, Hannah M (2010) A bootstrap approach to assess parameter uncertainty in simple catchment models. *Environ Model Softw* 25:919–926
- Shen JC, Chang CH, Wu SJ, Hsu CT, Lien HC (2015) Real-time correction of water stage forecast using combination of forecasted errors by time series models and Kalman filter method. *Stoch Environ Res Risk Assess*. doi:[10.1007/s00477-015-1074-9](https://doi.org/10.1007/s00477-015-1074-9)
- Sivakumar B (2008a) The more things change, the more they stay the same: the state of hydrologic modelling. *Hydrol Process* 22:4333–4337
- Sivakumar B (2008b) Dominant processes concept, model simplification and classification framework in catchment hydrology. *Stoch Environ Res Risk Assess* 22:737–748
- Tongal H, Berndtsson R (2016) Impact of complexity on daily and multi-step forecasting of streamflow with chaotic, stochastic, and black-box models. *Stoch Environ Res Risk Assess*. doi:[10.1007/s00477-016-1236-4](https://doi.org/10.1007/s00477-016-1236-4)
- Wagner T, Gupta HV (2005) Model identification for hydrological forecasting under uncertainty. *Stoch Environ Res Risk Assess* 19:378–387
- Wagner T, Wheeler HS, Gupta HV (2004) *Rainfall–runoff modelling in gauged and ungauged catchments*. Imperial College Press, London
- Wang YC, Yu PS, Yang TC (2010) Comparison of genetic algorithms and shuffled complex evolution approach for calibrating distributed rainfall–runoff model. *Hydrol Process* 24:1015–1026
- Wang Y, Guo S, Chen H, Zhou Y (2014) Comparative study of monthly inflow prediction methods for the Three Gorges Reservoir. *Stoch Environ Res Risk Assess* 28:555–570
- Wei CC (2014) Simulation of operational typhoon rainfall nowcasting using radar reflectivity combined with meteorological data. *J Geophys Res Atmos* 119:6578–6595. doi:[10.1002/2014JD021488](https://doi.org/10.1002/2014JD021488)
- Willmott CJ (1981) On the validation of models. *Phys Geogr* 2:184–194
- Wu CL, Chau KW, Fan C (2010) Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. *J Hydrol* 389:146–167
- Wu SJ, Lien HC, Chang CH, Shen JC (2012) Real-time correction of water stage forecast during rainstorm events using combination of forecast errors. *Stoch Environ Res Risk Assess* 26:519–531
- Yaseen ZM, El-shafie A, Jaafar O, Afan HA, Sayl KN (2015) Artificial intelligence based models for stream-flow forecasting: 2000–2015. *J Hydrol* 530:829–844
- Yen H, Hoque Y, Harmel RD, Jeong J (2015) The impact of considering uncertainty in measured calibration/validation data during auto-calibration of hydrologic and water quality models. *Stoch Environ Res Risk Assess* 29:1891–1901
- Yu B, Sombatpanit S, Rose CW, Ciesiolka CAA, Coughlan KJ (2000) Characteristics and modeling of runoff hydrographs for different tillage treatments. *Soil Sci Soc Am J* 64:1763–1770