Front. Earth Sci. 2020, 14(1): 188–200
https://doi.org/10.1007/s11707-019-0773-9

RESEARCH ARTICLE

# Probabilistic forecasting based on ensemble forecasts and EMOS method for TGR inflow

**Yixuan ZHONG[1,2], Shenglian GUO (✉)[1], Feng XIONG[1], Dedi LIU[1], Huanhuan BA[1], Xushu WU[1]**

1 State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China
2 China Water Resources Pearl River Planning Surveying & Designing Co, Ltd., Guangzhou 510610, China



**Abstract** Probabilistic inflow forecasts can quantify the uncertainty involved in the forecasting process and provide useful risk information for reservoir management. This study proposed a probabilistic inflow forecasting scheme for the Three Gorges Reservoir (TGR) at 1–3 d lead times. The post-processing method Ensemble Model Output Statistics (EMOS) is used to derive probabilistic inflow forecasts from ensemble inflow forecasts. Considering the inherent skew feature of the inflow series, lognormal and gamma distributions are used as EMOS predictive distributions in addition to conventional normal distribution. Results show that TGR's ensemble inflow forecasts at 1–3 d lead times perform well with high model efficiency and small mean absolute error. Underestimation of forecasting uncertainty is observed for the raw ensemble inflow forecasts with biased probability integral transform (PIT) histograms. The three EMOS probabilistic forecasts outperform the raw ensemble forecasts in terms of both deterministic and probabilistic performance at 1–3 d lead times. The EMOS results are more reliable with much flatter PIT histograms, coverage rates approximate to the nominal coverage 89.47% and satisfactory sharpness. Results also show that EMOS with gamma distribution is superior to normal and lognormal distributions. This research can provide reliable probabilistic inflow forecasts without much variation of TGR's operational inflow forecasting procedure.

**Keywords** ensemble forecast, probabilistic forecast, numeric weather prediction, EMOS, Three Gorges Reservoir

## 1 Introduction

Real-time flood forecast plays an important role in water resources management, flood control, and environmental protection activities (Chen et al., 2015). The World Meteorological Organization defines flood forecasting as an important tool for reducing vulnerabilities and flood risk and form an important ingredient of the strategy to "live with floods," thereby contributing to national sustainable development (WMO, 2010). Considering the uncertainty involved in flood forecasting processes, people nowadays prefer probabilistic forecasts, which take the form of a predictive probability density function (PDF) rather than the conventional deterministic or single-value forecast (Gneiting and Katzfuss, 2014; Huang et al., 2018; Liu et al., 2018). Previous literature has shown that probabilistic forecasts have the following advantages (Bourdin et al., 2014; Hardy et al., 2016; Liu et al., 2016; Todini, 2017): 1) the PDF provided by probabilistic flood forecasts enables the decision makers to quantify uncertainty and therefore better trade off risks and benefits; 2) probabilistic flood forecasts usually have longer lead time and thus provide more timely flood information; and 3) probabilistic flood forecasts usually outperform deterministic flood forecasts on forecasting skills. Therefore, probabilistic forecasts can better serve stakeholders and reservoir management with the ability to quantify the prediction uncertainty, and have thus become an essential ingredient of optimal decision making.

Probabilistic forecasts are usually generated by post-processing ensemble forecasts (Cloke and Pappenberger, 2009; Gneiting and Katzfuss, 2014). Ensemble forecasts are regarded as random samples of the future status. An instinct innovation is to use the ensemble forecasts to estimate the parameters of the PDFs. However, the ensemble forecasts are more or less biased and dispersive, statistical post-processing methods are therefore necessary for generating calibrated probabilistic forecasts (Bröcker

Received December 24, 2018; accepted June 10, 2019

E-mail: slguo@whu.edu.cn

and Smith, 2007; Wilks and Hamill, 2007; Bourdin et al., 2014; Zhong et al., 2018a). Two statistical post-processing methods are most popular nowadays, i.e., Bayesian model averaging (BMA) (Raftery et al., 2005) and ensemble model output statistics (EMOS) (Gneiting et al., 2005). The BMA method is a weighted average of the individual posterior conditional PDFs between the ensemble members and the observations, while the EMOS method post-processes the ensemble forecasts with a single parametric distribution. Previous literature shows that BMA and EMOS have comparable performance in post-processing hydro-meteorological variables, such as wind speed, temperature, atmospheric pressure, precipitation, flood, etc. (Raftery et al., 2005; Duan et al., 2007; Sloughter et al., 2007; Liu and Xie, 2014; Baran and Nemoda, 2016). In this study we employ the EMOS method for ensemble post-processing, which is flexible and has successful applications in many relevant studies (Gneiting and Raftery, 2005; Bourdin et al., 2014; Hemri et al., 2015).

The objective of probabilistic forecasting is to maximize the sharpness of predictive PDFs subject to calibration or reliability (Gneiting et al., 2007). Sharpness refers to the concentration of the predictive distributions in absolute terms and is a property exclusive to the forecasts and calibration or reliability refers to the statistical compatibility of probabilistic forecasts and observations (Gneiting and Katzfuss, 2014). Accurate ensemble forecasts which sample different uncertainty sources are needed before applying a post-processing method. Zhao et al. (2012) generated ensemble flood forecasts with different numeric weather prediction (NWP) products and post-processed the ensemble forecasts with BMA method. Naiafi and Moradkhani (2014) improved the BMA method with Copula functions and applied it to multi-model flood simulations to generate probabilistic forecasts. Hemri et al. (2015) used EMOS method to post-process hydrologic ensemble forecasts that consider input uncertainty and applied empirical copula coupling to reconstruct the temporal structure of the multistep probabilistic forecasts. Wu et al. (2016) conducted atmospheric-hydrological modeling based on ensemble precipitation forecasting and obtained ensemble hydrographs for a flood event. Najafi and Moradkhani (2016) tested multi-model ensemble averaging techniques over several basins and revealed that the BMA expectations performed best. Previous literature shows that raw ensemble forecasts are usually probabilistically biased and post-processing is needed. Moreover, the predictive PDFs describe the inherent nature of the prediction variables and should thus be selected specifically for different variables. For instance, Gaussian distribution for temperature and sea level pressure (Gneiting and Rafrery, 2005; Liu and Xie, 2013), lognormal and generalized extreme value distribution for wind speed (Lerch and Thorarinsdottir, 2013; Baran and Lerch, 2015), lognormal for reservoir inflow (Bourdin et al., 2014), mixed logistic-Gamma distribution for precipitation (Sloughter et al., 2007). Thus, the predictive distribution for generating probabilistic inflow forecasts should be carefully selected.

The main purpose of this research is to generate reliable probabilistic forecasts from ensemble forecasts for reservoir inflow. Different distributions are used for conducting EMOS post-processing. The EMOS probabilistic forecasts are evaluated and compared with the raw ensemble forecasts in terms of both deterministic and probabilistic performance. Three Gorges Reservoir (TGR) in China is selected as a case study to illustrate our approach. The remainder of this paper is organized as follows. Section 2 displays the basic information of the study area and data used. Section 3 explains methodologies and evaluation criteria. Section 4 presents and discusses the main results of this study. Finally, conclusions are illustrated in Section 5.

## 2 Study area and data

### 2.1 Study area

Three Gorges Reservoir (TGR) is the largest hydroelectric project located on the world's third longest river, the Yangtze River, which protects millions of people from flood disasters and produces about 100 billion kW/h hydropower annually (Zhong et al., 2018b). TGR has comprehensive benefits of flood control, hydropower generation, navigation, environmental protection, among other benefits. The drainage area of TGR is about 1 million $km^2$ and its interval basin is about 59100 $km^2$, which is displayed in Fig. 1. The inflow of TGR interval basin consists of three parts: 1) the mainstream inflow gauged by Cuntan hydrologic station, 2) the tributary inflow gauged by Wulong hydrologic station and 3) the rainfall-runoff of the uncontrolled interval basin, which accounts for about 6% of the TGR basin. The operational inflow forecasting system of TGR has been run since 2003 by the Changjiang Water Resources Commission (CWRC) and can provide short-term deterministic forecasts of TGR inflow as well as flood forecasts of upstream hydrologic stations.

### 2.2 Hydrological data

To implement inflow forecasting of TGR with 1–3 d lead times, the following hydrological data are collected from CWRC, including 1) gauged daily discharges at Cuntan and Wulong stations; 2) 1–2 d daily flood forecasts at Cuntan and Wulong stations; 3) mean daily precipitation records and 4) precipitation predictions over TGR's interval basin. Inflow observations of TGR are also collected for model calibration purposes. The period of the data sets is from 2010 to 2015.

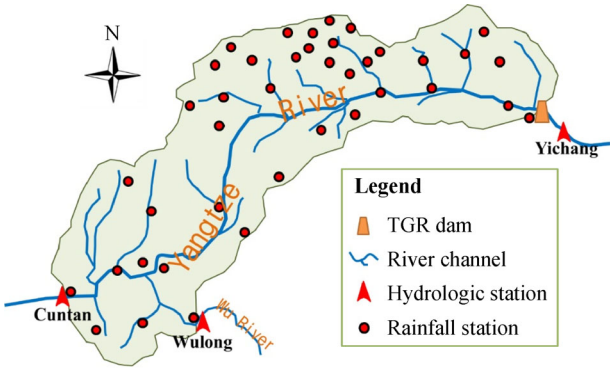


**Fig. 1** Sketch map of the TGR interval basin.

### 2.3 NWP data

NWP data are necessary for accurate inflow forecasting (Khan et al., 2015). The NWP data used in this study are obtained from the European Center for Medium-range Weather Forecast (ECMWF), the National Center for Environmental Prediction (NCEP) and the China Meteorological Administration (CMA) of The Interactive Grand Global Ensemble (TIGGE, WMO, 2005) project, which can be downloaded from its official website. The three NWP suites each has 51, 21, and 15 members. The raw NWP data are grid-based. To fit the hydrologic models, NWP data are first downscaled to basin scale by arithmetically averaging the NWP values at all grids within TGR's interval basin following Duan et al. (2007), and then averaged within each data suite (ECMWF, CMA and NCEP). Eventually, three-member NWP daily precipitation predictions at 1–3 d lead times are used as hydrologic model input to obtain forecast inflows of the interval basin. The NWP data are not post-processed before driving the hydrologic models. As revealed by Mascaro et al. (2011), although further correction of end forecast is required, only adjustment of the final hydrologic forecast is strictly necessary from the operation point of view, which can be implemented by the EMOS method.

## 3 Methodologies

The flowchart of establishing a probabilistic inflow forecasting scheme based on ensemble forecasts is displayed in Fig. 2. The main methodologies employed are introduced as follows:

### 3.1 Ensemble inflow forecasting scheme of TGR

According to the operational inflow forecasting procedure of TGR, the floods in the mainstream gauged by the Cuntan hydrologic station and the Wu River gauged by the Wulong hydrologic station are routed to the reservoir using the Muskingum-Cunge propagation method (Cunge, 1969). The areal mean precipitation is transformed into runoff with hydrologic models to estimate the interval basin inflow. Then TGR inflow is estimated with the summation of the three flow components.

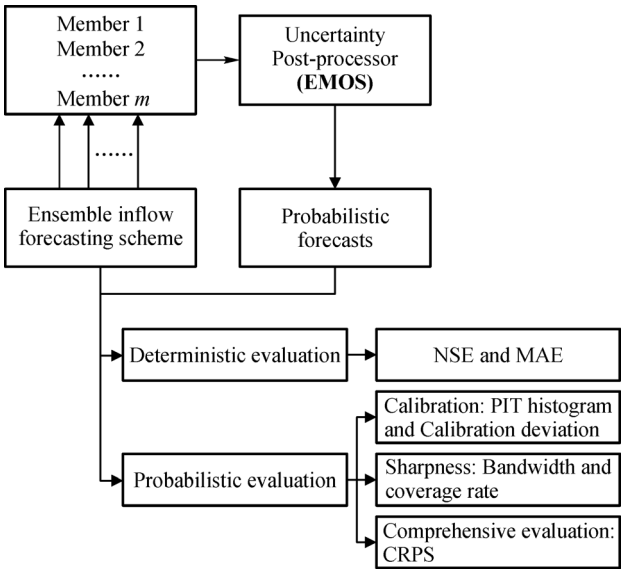


**Fig. 2** Flowchart of probabilistic inflow forecasting scheme based on ensemble NWP and EMOS method.

To sample the forecasting uncertainty, an ensemble inflow forecasting scheme is developed on the basis of TGR's operational inflow forecasting system in this study. Three major sources of hydrologic uncertainty are considered (Emam et al., 2018), i.e., 1) input data, including the NWP data errors and upstream inflow forecasting errors; 2) model structure, which is subject to the inability of a single model to coincide with all kinds of hydrologic processes and 3) model parameter, which comes from the equifinality for different parameters and estimation methods. The outcomes of this ensemble forecasting scheme are 18 inflow forecasts of TGR (3 inputs × 2 models × 3 parameter sets = 18 members) at 1–3 d lead times. The schematic of TGR ensemble forecasting scheme is displayed in Fig. 3.

#### 3.1.1 Hydrologic models

Two lumped conceptual models, i.e., the Xinanjiang (XAJ) model (Zhao, 1992) and GR4J model (Perrin et al., 2003) are used to model the interval basin inflow of TGR. Inflows of the Cuntan and Wulong hydrologic stations are routed using the Muskingum-Cunge method. Thus, the inflow forecast of TGR at time $t$ can be expressed as:

$$Q_{\text{inflow}}(t) = Q_{\text{interval}}(t) + C_1 Q_{ct}(t-1) + C_2 Q_{wl}(t-1), \quad (1)$$

where $Q_{\text{inflow}}(t)$, $Q_{ct}(t)$ $Q_{wl}(t)$ and $Q_{\text{interval}}(t)$ denote the

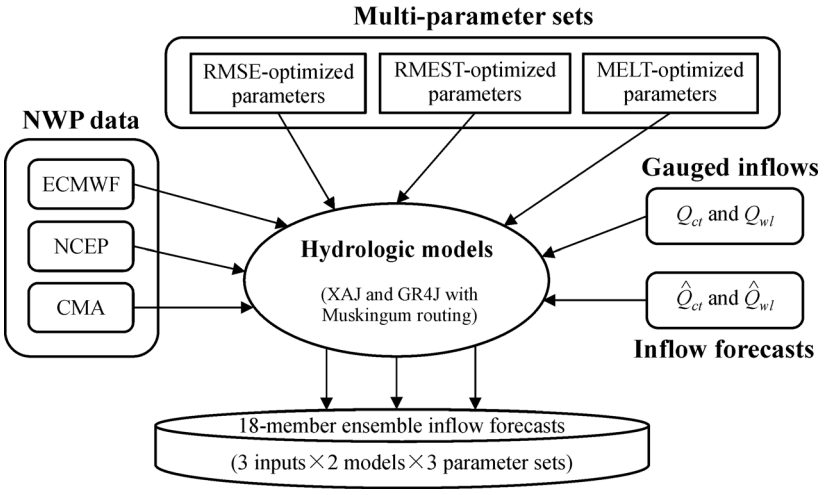


**Fig. 3** Schematic of the TGR ensemble inflow forecasting scheme

inflow of TGR, the discharges at Cuntan and Wulong stations and the inflow of TGR interval basin at time $t$, respectively; $C_1$ and $C_2$ are the propagation parameters for $Q_{ct}$ and $Q_{wl}$, respectively. Since this research focuses on real-time inflow forecasting, flood forecasts at the Cuntan and Wulong hydrologic stations are needed for lead times longer than 1 d and the interval basin flow is estimated by coupling NWP data with the calibrated hydrologic models.

The XAJ model is a lumped rainfall-runoff model for streamflow modeling in humid and semi-humid regions based on the concept of the saturation excess runoff mechanism (Dunne, 1978). The XAJ model performs well in humid and semi-humid regions with many successful applications (Hu et al., 2005; Tian et al., 2013; Jiang et al., 2014; Lin et al., 2014; Jie et al., 2018). The state-of-the-art version of XAJ model is the three-component one, which has 14 uniform parameters. A detailed illustration of the XAJ model is beyond the scope of this research; readers can refer to Zhao (1992) for more information.

The GR4J model is a lumped rainfall-runoff model with four parameters. Conceptually, the river basin can be represented by two reservoirs, i.e. a soil reservoir and a routing reservoir. After subtracting potential evapotranspiration from the precipitation, the net precipitation is divided into two portions. One portion goes into the soil reservoir, which will later be drained by either evapotranspiration or percolation toward deep flow. The other portion is routed directly to the outlet. The two flow components meet together and 90% of which is routed by a unit hydrograph and then a nonlinear routing store. The last 10% is routed just by another unit hydrograph. The total runoff is eventually derived by gathering these two parts. Previous research proved that the GR4J model is as effective as more complex models with case studies of 429 basins all over the world (Perrin et al., 2001).

### 3.1.2 Evaluation metrics

To consider and emphasize different parts of inflow simulations (Oudin et al., 2006; Duan et al., 2007; Bourdin et al., 2014), three evaluation metrics are selected for model parameter calibration.

1) Root mean square error (*RMSE*)

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Q_{\text{obs}}(t) - Q_{\text{sim}}(t)\right)^2}, \qquad (2)$$

where $Q_{\text{obs}}(t)$ and $Q_{\text{sim}}(t)$ denote the observed and simulated inflows at time $t$, respectively; $n$ is the data length. This metric is sensitive to the differences between the observation and simulation, and is considered as an overall evaluation metric. A lower *RMSE* value indicates a better simulation performance.

2) Root mean error of square transformed (*RMEST*)

$$RMEST = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Q_{\text{obs}}^2(t) - Q_{\text{sim}}^2(t)\right)^2}. \qquad (3)$$

By transforming the observations and simulations into square form, this metric puts great emphasis on the peak flow discharges.

3) Mean error of logarithmic transformed (*MELT*)

$$MELT = \frac{1}{n}\sum_{t=1}^{n}\left(\ln Q_{\text{obs}}(t) - \ln Q_{\text{sim}}(t)\right)^2. \qquad (4)$$

A small MELT value emphasizes good fit of low flow discharges.

Genetic Algorithm (GA, Goldberg, 1989), which is robust and reliable even when there are many model parameters (Wu and Chau, 2006; Parasuraman and Elshorbagy, 2007; Kang et al., 2017), is used to calibrate model parameters.

### 3.2 Ensemble Model Output Statistics (EMOS)

The EMOS method proposed by Gneiting and Raftery (2005) is used to post-process the ensemble inflow forecasts, in which the normal distribution is assumed as the predictive PDF. The probabilistic forecasts can be expressed as follows.

$$g(y|x_1,x_2,...,x_m) = N(a + \sum_{i=1}^{m} b_i x_i, c + dS^2), \quad (5)$$

where $N(\cdot)$ denotes normal PDF; $a + \sum_{i=1}^{m} b_i x_i$ is the affine function of the $m$ ensemble forecasts ($[x_1,x_2,...,x_m]$); $c + dS^2$ is the affine function of the ensemble variance $S^2$; $a$, $b_i$, $c$ and $d$ are the EMOS parameters.

By adopting other distributions (e.g. lognormal, P-III, gamma) as predictive distribution, the EMOS method can be generalized and applied to different variables (Wilks and Hamill, 2007). In addition to the normal distribution, lognormal and gamma distributions are also employed as EMOS predictive distributions in this study. The reason for choosing these two non-Gaussian distributions is that they have a long history as well as many successful cases in hydrologic application (Fernandez and Salas, 1986; Lewis et al., 2000; Yue et al., 2001; Steinschneider and Brown, 2011; Bourdin et al., 2014; Xiong et al., 2018). For instance, Bourdin et al. (2014) used Gaussian and log-normal distributions to conduct EMOS post-processing and obtained reliable probabilistic inflow forecasts. Also, the EMOS based on log-normal and gamma distributions are left-truncated at zero like the one proposed by Thorarinsdottir and Gneiting (2010), which can ensure the post-processed inflow forecasts are positive.

The lognormal PDF is expressed as:

$$g(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right), \quad (6)$$

where $\mu$ and $\sigma$ denote the mean and standard deviation of the log-transformed $x$.

The gamma PDF is expressed as:

$$g_{k,\theta}(y) = \frac{x^{k-1}\exp\left(-\frac{x}{\theta}\right)}{\theta^k\Gamma(k)},\ x>0, \quad (7)$$

where $\Gamma(\,)$ denotes the gamma function; $k$ and $\theta$ are the distribution parameters, which are related with the population mean $\mu$ and standard deviation $\sigma$ by:

$$k = \frac{\mu^2}{\sigma^2}, \quad (8)$$

$$\theta = \frac{\sigma^2}{\mu}. \quad (9)$$

The EMOS method using ensemble forecasts as inputs, the $\mu$ and $\sigma$ can be estimated by the following equations:

$$\mu = a + \sum_{i=1}^{m} b_i x_i, \quad (10)$$

$$\sigma^2 = c + dS^2, \quad (11)$$

where $S^2$ denotes the variance of the ensemble forecasts, $a$, $b_i$, $c$ and $d$ are the EMOS parameters.

The EMOS parameters are optimized by the genetic algorithm (GA) method with minimizing the continuous ranked probability score (*CRPS*) as objective function during calibration period. The EMOS parameters are adaptively updated with sliding window method, in which the previous $n_w$ data samples are utilized for parameter optimization for each time step as illustrated by Fig. 4. The underlying assumption of sliding window method is that the most recent samples can best represent the forecasting feature at present. When using sliding window method, the window length $n_w$ must be carefully specified. If the value of $n_w$ is too large, then the model may be overfitting; on the contrary, the model may be underfitting. The *CRPS* values of EMOS probabilistic forecasts with different $n_w$ from 20 to 120 are calculated and the $n_w$ corresponding to the smallest *CRPS* value is selected in this study.

### 3.3 Evaluation Criteria

Nash-Sutcliff coefficient of efficiency (*NSE*, also defined as model efficiency) and Mean Absolute Error (*MAE*) are selected to evaluate the performance of deterministic forecasts (Nash and Sutcliffe, 1970; Lan et al., 2018).

$$NSE = 1 - \frac{\sum_{t=1}^{n}\left(Q_{\text{obs}}(t) - Q_{\text{sim}}(t)\right)^2}{\sum_{t=1}^{n}\left(Q_{\text{obs}}(t) - \overline{Q}_{\text{obs}}\right)^2}, \quad (12)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Q_{\text{obs}}(t) - Q_{\text{sim}}(t)|, \quad (13)$$

where $\overline{Q}_{\text{obs}}$ denotes the mean observed discharge of the $n$ samples.

Probability integral transform (*PIT*) histogram, calibration deviation (*CD*), bandwidth (*BD*), and coverage rate (*CR*) are selected to evaluate reliability and sharpness of probabilistic forecasts. *CRPS* is used to make comprehensive evaluation.

Perfect probabilistic forecasts are expected to have a flat *PIT* histogram. The *PIT* value for the $t$th sample is:

$$PIT_t = G_t(O_t), \quad (14)$$

where $O_t$ denotes the $t$th observation; $G_t(\,)$ denotes the probabilistic cumulative distribution function (CDF) at time $t$, which is the integration of Eqs. (6)–(7).

The PIT values are then assigned to different bins and the frequency of each bin is calculated. The number of *PIT* bins is usually selected arbitrarily from 10 to 20. According to the shape of the *PIT* histogram, the problems

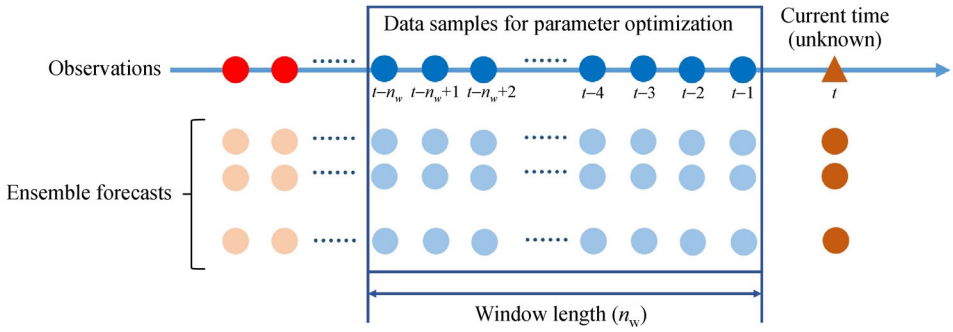


**Fig. 4** Illustration of sliding window method for adaptive parameter optimization.

of the probabilistic forecasts can be diagnosed. If the *PIT* histogram is U-shaped, it usually indicates the probabilistic forecasts have inadequate spread or under-dispersion. On the contrary, if the *PIT* histogram is humpback-shaped, then the probabilistic forecasts are recognized as over-dispersive. However, it is noteworthy that a flat *PIT* histogram is not a sufficient condition for perfect calibration or reliability, since a combination of negatively and positively biased forecasting distributions can also yield a flat *PIT* histogram while being unreliable (Hamill, 2001). The *CD* derived from the *PIT* values can give a more quantitative evaluation of reliability:

$$CD = \sqrt{\frac{1}{h}\sum_{i=1}^{h}\left(\mathrm{bin}_i - \frac{1}{h}\right)^2}, \quad (15)$$

where $\mathrm{bin}_i$ is the bin frequency of the $i$th bin; $h$ is the number of the bins. A small *CD* value is preferred, which means the deviation from a flat *PIT* histogram is negligible.

*BD* assesses the sharpness of probabilistic forecasts. The average *BD* for the probabilistic forecasts can be computed as:

$$BD = \frac{1}{n}\sum_{t=1}^{n}\left(G_t^{-1}(1-\alpha/2) - G_t^{-1}(\alpha/2)\right), \quad (16)$$

where $G_t^{-1}(\,)$ denotes the inverse CDF of the $t$th sample; $\alpha$ is the confidence degree. The mathematical meaning of *BD* is apparent that this index represents the uncertainty range of the probabilistic forecast. Since probabilistic forecasts with similar *BD* values may have significant differences in capturing the observations, coverage rate (*CR*) can help further differentiate the results:

$$CR = \frac{1}{n}\sum_{t=1}^{n}\left(G_t^{-1}(1-\alpha/2) \leqslant O_t \leqslant G_t^{-1}(\alpha/2)\right), \quad (17)$$

where $(\,)$ is 1 when the condition is satisfied and otherwise 0. *CR* is a positive oriented index with a range of [0, 1]. Theoretically, the *CR* value should be close to the confidence degree $\alpha$. Only in this context, the smaller the *BD* is, the better.

*CRPS* characterizes both reliability and sharpness of probabilistic forecasts (Gneiting et al., 2007):

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left[G_t(O_t) - H(x - O_t)\right]^2 \mathrm{d}x, \quad (18)$$

where $H$ is the Heaviside function. *CRPS* ranges in $[0, \infty)$ with zero as the perfect value. A lower *CRPS* value indicates a better probabilistic performance. Note that for deterministic forecasts, the *CRPS* reduces to *MAE* and make it possible to compare probabilistic forecasts with deterministic forecasts.

## 4 Results and discussion

### 4.1 Ensemble inflow forecasting performance

The performance of inflow simulations with different hydrologic models and objective functions during calibration period (2010–2013) is shown in Table 1. Results show that the six inflow forecasting schemes yield comparable performances with all the *NSE* values larger than 0.99. Since the Cuntan and Wulong hydrologic stations control about 94% of the basin area, and the TGR is a type of river channel reservoir, the satisfactory modeling performance of inflow is expected and reasonable. The *RMSE*, *RMEST* and *MELT* values are different from each other, showing a good ability for sampling uncertainty. Duan et al. (2007) and Bourdin et al. (2014) have likewise conducted multi-model and multi-objective function approaches to generate ensemble flood forecasts.

Subsequently, the NWP data from the three TIGGE center are used as forces for the six inflow forecasting schemes to generate ensemble inflow forecasts at 1–3 d lead times. Subject to the available flood forecasts of the Cuntan and Wulong stations, real-time ensemble inflow forecasts of TGR range from May, 11st, 2012 to December, 31st, 2015 (1330 days in total). The evaluation metrics for the ensemble flow forecasts are shown in Fig. 5 with Box-Whisker plots. Generally speaking, the ensemble inflow

**Table 1** Performances of the inflow forecasting schemes during calibration period

| Hydrologic model | | XAJ | | | GR4J | | |
|---|---|---|---|---|---|---|---|
| Optimization function | | *RMSE* | *RMEST* | *MELT* | *RMSE* | *RMEST* | *MELT* |
| Evaluation metrics | RMSE | 935.6 | 944.4 | 945.0 | **888.2** | 893.8 | 937.4 |
| | RMEST | $3.00\times10^6$ | $\mathbf{1.31\times10^6}$ | $3.56\times10^6$ | $4.53\times10^6$ | $3.00\times10^6$ | $4.63\times10^6$ |
| | MELT | 0.0046 | 0.0048 | **0.0045** | 0.0049 | 0.0049 | 0.0046 |
| | NSE | 0.9909 | 0.9907 | 0.9907 | **0.9918** | 0.9917 | 0.9909 |

Note: * The **BOLD** values indicate the best results for each evaluation metrics.

forecasts yield satisfactory results with *NSE* medians 0.98, 0.94 and 0.89 at 1–3 d lead times, respectively. The *RMSE*, *RMEST,* and *MELT* values are also satisfactory while slightly increasing with increase of lead time. Figure 5 reveals that forecasting uncertainty increases with the lead time as indicated by the more dispersive metric values.

### 4.2 Optimization of window length $n_w$

The EMOS parameters are updated adaptively via a sliding window training method, of which the window length $n_w$ is a vital parameter. If $n_w$ is too small, the estimated model parameters may converge prematurely. On the other hand, the parameters may be overfitting with a large $n_w$ value. Figure 6 shows the relationship of *CRPS* values with different $n_w$ values. It can be seen from Fig. 6 that for different EMOS schemes, the *CRPS* values generally decrease as $n_w$ increases. The curves turned to be flat after 70–80 days for the three lead times. Results indicate that when $n_w$ is larger than 80 days, the *CRPS* values have stabilized. Thus, the $n_w$ is set to be 80 days hereafter for all three EMOS methods at 1-3d lead times.

### 4.3 Evaluation of deterministic forecasts

The EMOS methods with normal (NOR-EMOS), lognormal (LN-EMOS) and gamma distributions (GM-EMOS) are used to post-process the 18-member ensemble inflow forecasts of TGR, respectively. Deterministic inflow forecasts are defined as arithmetic mean values of the raw ensemble forecasts and expectations of the EMOS PDFs, respectively. Table 2 compares the deterministic evaluation results of the raw ensemble inflow forecasts with the EMOS probabilistic inflow forecasts in terms of *NSE* and *MAE*. After EMOS post-processing, the *MAE* at 1–3 d lead times has been efficiently reduced, while the improvement is slight in terms of *NSE*.

The reductions of deterministic bias are mainly due to that the EMOS method ensures ensemble members that have better performance be allocated with larger weights. As can be seen from Table 2, the *MAE* and *NSE* values of the three EMOS deterministic inflow are quite close to each other. Results indicate that different predictive distributions can obtain similar deterministic performance. The explanation is that deterministic performance of

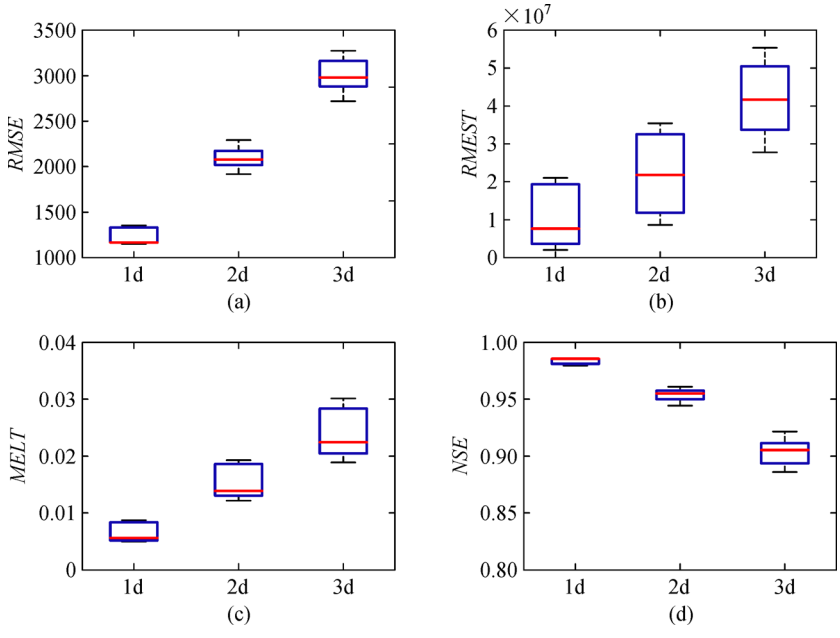


**Fig. 5** Box-whisker plots for the evaluation metrics of TGR ensemble inflow forecasts at 1–3 d lead time.

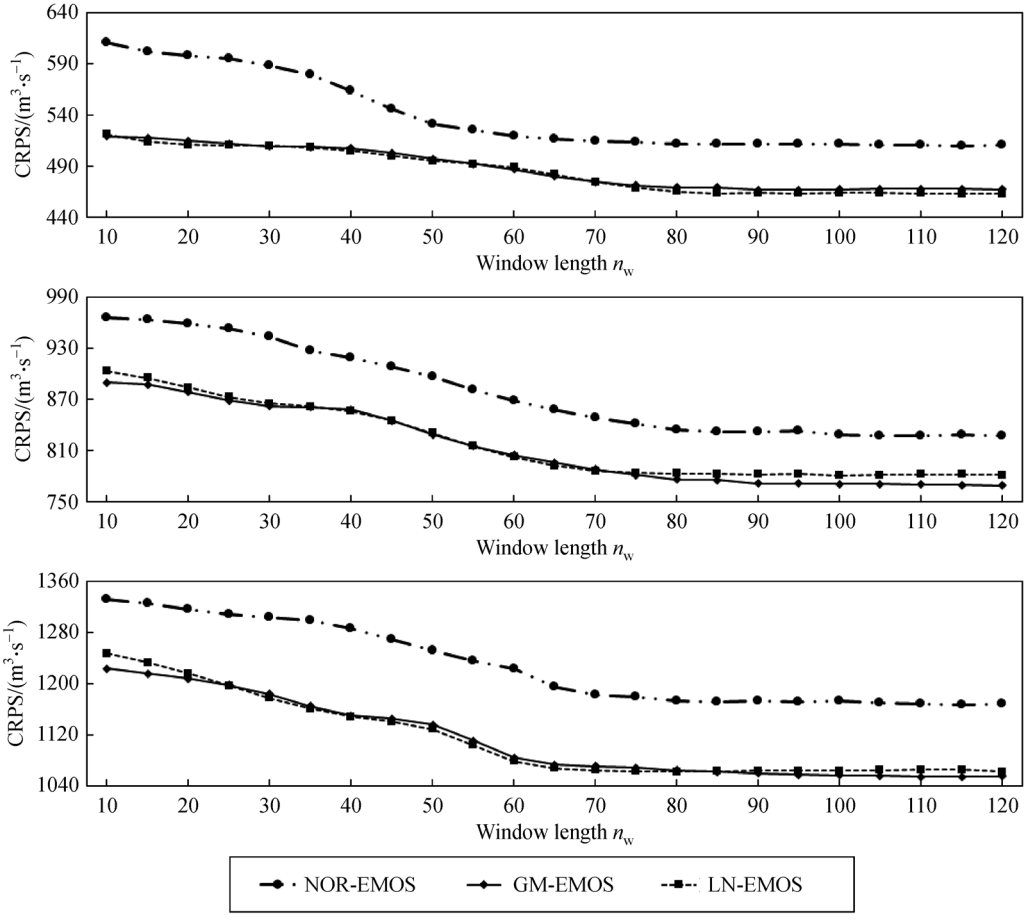


**Fig. 6** CRPS values with different window length $n_w$ for 1–3 d EMOS probabilistic inflow forecasts

**Table 2** Comparison of raw ensemble and EMOS probabilistic forecasts at 1–3 d lead time in terms of deterministic and probabilistic performance

| Method | Lead time | Deterministic metrics | | Probabilistic metrics | | |
|---|---|---|---|---|---|---|
| | | MAE ($m^3/s$) | *NSE* | CRPS ($m^3/s$) | BD ($m^3/s$) | CR*(%) |
| Raw ensemble | 1d | 675 | 0.98 | 675 | 837 | 34.44 |
| | 2d | 1126 | 0.94 | 1126 | 1002 | 27.59 |
| | 3d | 1597 | 0.89 | 1597 | 1105 | 26.77 |
| NOR-EMOS | 1d | 628 | 0.98 | 511 | 3423 | 86.58 |
| | 2d | 1036 | 0.96 | 834 | 4858 | 85.20 |
| | 3d | 1407 | 0.90 | 1174 | 6368 | 83.83 |
| GM-EMOS | 1d | 628 | 0.99 | 469 | 3064 | 89.68 |
| | 2d | 1048 | 0.96 | 776 | 4276 | 86.96 |
| | 3d | 1399 | 0.91 | 1065 | 5160 | 86.68 |
| LN-EMOS | 1d | 624 | 0.99 | 465 | 3341 | 90.56 |
| | 2d | 1034 | 0.95 | 783 | 4879 | 88.64 |
| | 3d | 1388 | 0.91 | 1063 | 6908 | 89.61 |

Note: * The nominal coverage is $(18-1)/(18+1)=89.47\%$ according to the size of ensemble forecasts.

EMOS probabilistic forecasts is only attributed to the estimation of distribution expectations, which is essentially first moment of the sample series and its estimation is robust and easy, thus the selection of predictive distribution seems of no consequence.

### 4.4 Evaluation of probabilistic forecasts

The PIT histograms are used to illustrate the probabilistic calibration or reliability as shown in Fig. 7. As stated previously, reliable probabilistic forecasts should have a flat PIT histogram. The raw ensemble inflow forecasts show bad reliability with the inflow observation concentrates in the first and last bins. The badly-shaped PIT histograms indicate that the raw ensemble is biased and have severely underestimated the forecasting uncertainty, which indicates potential risk. The concentration of PIT values in high percentile indicates that the risk of flood occurrence is significantly underestimated by the raw ensemble spreads, which may lead to wrong operation decisions. After EMOS post-processing, the PIT histograms are much flatter at 1–3 d lead times. Underdispersion of the raw ensemble inflow forecasts has been efficiently ameliorated by the EMOS method. NOR-EMOS results are less reliable than the other two EMOS results, indicating that skewed distributions can better describe the uncertainty spreads of inflow forecasts. This has clearly illustrated that selecting an appropriate predictive distribution can yield probabilistic forecasts with good calibration and vice versa. Despite the three EMOS results have approximate deterministic performance, the NOR-EMOS fails to produce reliable prob-

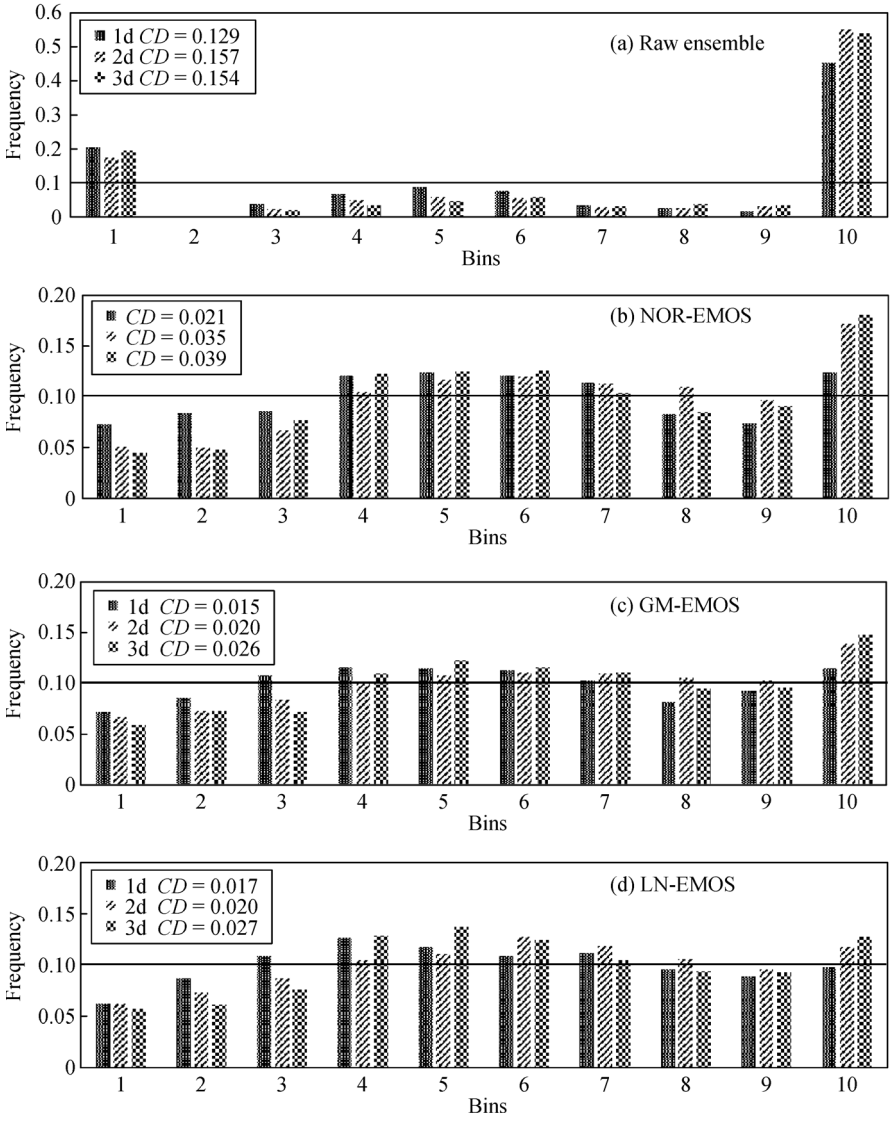


**Fig. 7** Calibration evaluation using Verification Ranked Histogram for (a) raw ensemble forecasts and PIT histograms for (b) NOR-EMOS, (c) LN-EMOS and (d) GM-EMOS probabilistic forecasts at 1–3 d lead times. The horizontal line indicates the ideal uniform distribution. The corresponding *CD* values are displayed in the legends.

abilistic forecasts since inflow series of TGR are skewed or non-Gaussian (Zhong et al., 2018b).

Table 2 displays the probabilistic evaluation results of the raw ensemble forecasts and the three EMOS probabilistic forecasts in terms of *BD*, *CR* and *CRPS*. Recall that sharpness is only meaningful when calibration is satisfied, the raw ensemble forecasts are regarded as the worst with *CR* values much smaller than the nominal coverage 89.47% even though they also have the smallest *BD* values or best sharpness. NOR-EMOS have *CR* values slightly smaller than the nominal coverage while the *CR* values of LN-EMOS and GM-EMOS are approximately 89.47%, indicating the latter two are more reliable in accordance with the PIT histograms and *CD* values. In fact, the *CR* values are stable with different lead times, but the *BD* values increase obviously with the lead time. This has added difficulty for stakeholders to make decision for long lead times since uncertainty is large. We argue that GM-EMOS have better probabilistic performance than LN-EMOS when both results are reliable. Similar conclusions can be drawn from the *CRPS* results, which show that the GM-EMOS and LN-EMOS have smaller *CRPS* values than the raw ensemble forecasts and NOR-EMOS.

To better distinguish the three EMOS results, their probabilistic inflow forecasts at 1–3 d lead times for annual maximum daily discharge in 2012 are shown in Fig. 8. The raw ensemble forecasts at 1 d lead time can well capture the inflow observation, but underestimate the inflow observation at 2 d and 3 d lead times. The EMOS probabilistic forecasts have wider spreads than the raw ensemble forecasts, while the inflow observation can be captured by the probabilistic intervals. Comparing the PDFs of the NOR-EMOS, LN-EMOS and GM-EMOS, it is found that forecasting PDFs of GM-EMOS have the

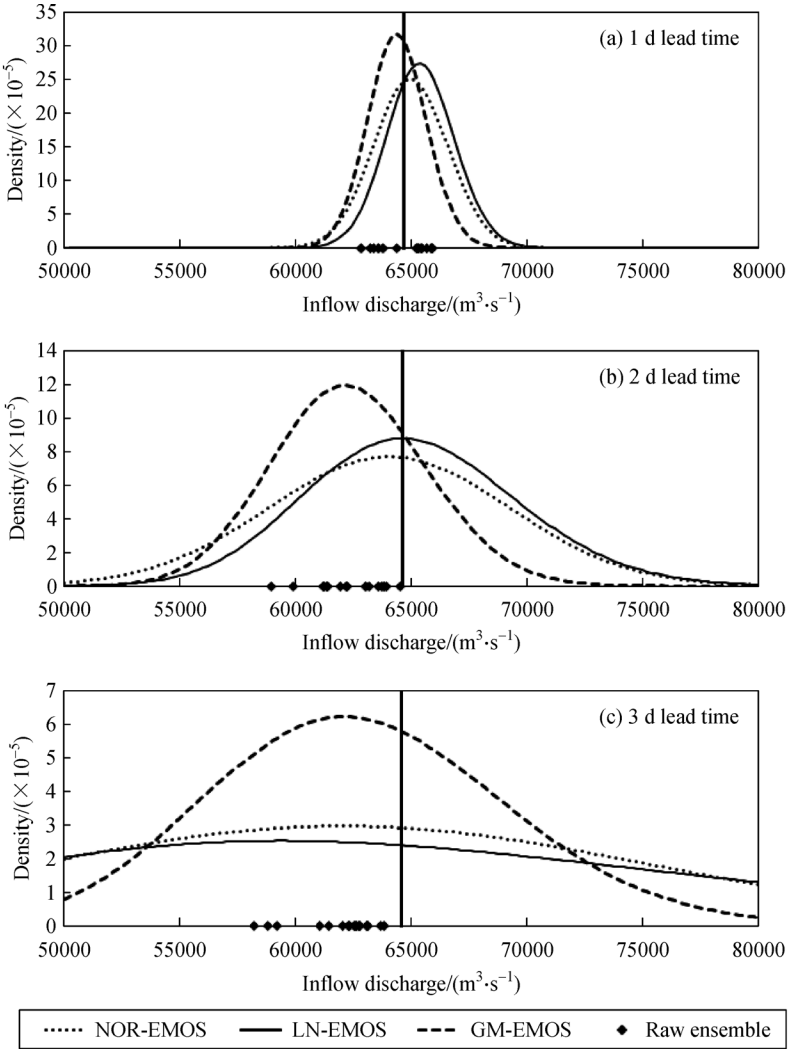


**Fig. 8** Probabilistic PDFs issued at 1–3 d lead time for daily inflow on 2012, July, 25th. The vertical solid lines indicate the inflow observation (64700 $m^3/s$).

narrowest spreads with the largest density allocated in the vicinity of the inflow observation. Given an overall consideration, the GM-EMOS method is recommended for ensemble post-processing of TGR inflow. It is also noted that the performance of probabilistic forecasts greatly relies on the accuracy of raw ensemble forecasts. As shown in Fig. 8, the PDFs of the three EMOS methods at 3 d lead time appear to have expectations smaller than the inflow observation since the raw ensemble forecasts are underestimated, which highlights the importance of improving ensemble forecasting skill.

### 4.5 Discussion

In this paper, we delivered a probabilistic inflow forecasting scheme of TGR with EMOS post-processing method based on ensemble forecasts. An 18-member ensemble inflow forecasting scheme based on TIGGE NWP data and different hydrologic models is developed and evaluated, which considers the forecasting uncertainty of input, model structure and parameter. Normal, lognormal, and gamma distributions are used to implement the EMOS post-processing. The raw ensemble inflow forecasts, NOR-EMOS, LN-EMOS and GM-EMOS probabilistic inflow forecasts are obtained and evaluated in terms of deterministic and probabilistic performance. Results indicate the raw ensemble forecasts of TGR inflow are severely under-dispersive and underestimate uncertainty. After EMOS post-processing, the probabilistic forecasts can improve both the performance of deterministic and probabilistic forecasting. The GM-EMOS is superior to NOR-EMOS and LN-EMOS and can generate the best PDFs for TGR inflow forecasting. Considering the complete inflow forecasting procedure of TGR involves over one hundred forecasting nodes and also great man-machine interaction, it will cost a lot of time, manpower, and money to develop a new forecasting scheme. This approach can fully utilize the outcomes of the TGR's operational inflow forecasting system, and change the deterministic forecasts to probabilistic forecasts easily.

Ensemble flow forecasting or probabilistic flow forecasting systems have been operated in many countries and districts (Cloke and Pappenberger, 2009; Laiolo et al., 2014). Arnal et al. (2016) carried out a risk-based decision-making game on the topic of flood protection mitigation, which is interesting and has contributed to communicate the probabilistic forecasters and the end-users. Todini (2017) made some progressive discussion on the application of probabilistic forecasts on flood early warning, adaptive reservoir management and the potential of combining risk-benefit with probabilistic forecasts.

It should be noted that achieving reliable probabilistic inflow forecasts is still far from end of the way. There is lack of operational case in China to date to provide risk information for stakeholders' decision making. The gap between forecasters and stakeholders on the knowledge of probabilistic forecasting should be filled since the conventional operation procedure based on deterministic forecast is still deep-rooted throughout the practical work. Therefore, continuous attention will be paid to improve the probabilistic inflow forecasting and serve water resources management of TGR.

## 5 Conclusions

The significance of probabilistic flow forecast for water resources management and flood control operation is widely acknowledged nowadays. A probabilistic forecasting scheme for reservoir inflow is developed based on ensemble forecasts and ensemble model output statistics (EMOS) method. The main conclusions of this study are summarized as follows:

1) Ensemble inflow forecasts of TGR at 1–3 d lead times perform well with the high *NSE* and small *MAE* values. They significantly underestimate forecasting uncertainty despite the major uncertainty sources sampled during ensemble generation process.

2) The EMOS probabilistic forecasts outperform the raw ensemble forecasts in terms of both deterministic and probabilistic evaluation criteria, and are more reliable with much flatter PIT histograms and *CR* values approximate to the nominal coverage.

3) LN-EMOS and GM-EMOS show better performance than NOR-EMOS, which emphasize the importance of using suitable predictive distributions for inflow series. GM-EMOS is recommended for TGR since it can generate calibrated or reliable probabilistic inflow forecasts with the least uncertainty among the three EMOS results as verified by the *BD* values.

4) The developed probabilistic inflow forecasting scheme can generate satisfactory forecasting PDFs with good reliability and sharpness and has potential value for practical application since it does not need much variation of the current inflow forecasting procedure of TGR.

**Acknowledgements** This study is supported by the National Key Research and Development Plan of China (No. 2016YFC0402206) and the National Natural Science Foundation of China (Grant Nos. 51879192, 91647106). Thanks are also given to CWRC for providing necessary data and the three anonymous reviewers' valuable suggestions to improve our manuscript.

## References

Arnal L, Ramos M H, de Perez E C, Cloke H L, Stephens E, Wetterhall F, van Andel S J, Pappenberger F (2016). Willingness-to-pay for a probabilistic flood forecast: a risk-based decision-making game. Hydrol Earth Syst Sci, 20(8): 3109–3128

Baran S, Lerch S (2015). Lognormal distribution based EMOS models for probabilistic wind speed forecasting. Q J R Meteorol Soc, 141 (691) 2289–2299

Baran S, Nemoda D (2016). Censored and shifted gamma distribution

based EMOS model for probabilistic quantitative precipitation forecasting. Environmetrics, 27(5): 280–292

Bourdin D R, Nipen T N, Stull R B (2014). Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system. Water Resour Res, 50(4): 3108–3130

Bröcker J, Smith L A (2007). Increasing the reliability of reliability diagrams. Weather Forecast, 22(3): 651–661

Chen L, Singh V P, Guo S, Zhou J, Zhang J (2015). Copula-based method for multisite monthly and daily streamflow simulation. J Hydrol (Amst), 528: 369–384

Cloke H L, Pappenberger F (2009). Ensemble flood forecasting: a review. J Hydrol (Amst), 375(3–4): 613–626

Cunge J A (1969). On the subject of a flood propagation computation method (Muskingum method). J Hydraul Res, 7(2): 205–230

Duan Q, Ajami N K, Gao X, Sorooshian S (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv Water Resour, 30(5): 1371–1386

Dunne T (1978). Field studies of hillslope flow processes. Hillslope hydrology, 227: 227–293

Emam A R, Kappas M, Fassnacht S, Linh N H K (2018). Uncertainty analysis of hydrological modeling in a tropical area using different algorithms. Front Earth Sci, 12(4): 661–671

Fernandez B, Salas J D (1986). Periodic gamma autoregressive processes for operational hydrology. Water Resour Res, 22(10): 1385–1396

Gneiting T, Balabdaoui F, Raftery A E (2007). Probabilistic forecasts, calibration and sharpness. J R Stat Soc, 69(2): 243–268

Gneiting T, Raftery A E (2005). Weather forecasting with ensemble methods. Science, 310(5746): 248–249

Gneiting T, Katzfuss M (2014). Probabilistic forecasting. J R Stat Soc, 1(1): 125–151

Goldberg D E (1989). Genetic algorithm in search, optimization, and machine learning. Addison Wesley: 2104–2116

Hamill T M (2001). Interpretation of rank histograms for verifying ensemble forecasts. Mon Weather Rev, 129(3): 550–560

Hardy J, Gourley J J, Kirstetter P E, Hong Y, Kong F, Flamig Z L (2016). A method for probabilistic flash flood forecasting. J Hydrol (Amst), 541: 480–494

Hemri S, Lisniak D, Klein B (2015). Multivariate postprocessing techniques for probabilistic hydrological forecasting. Water Resour Res, 51(9): 7436–7451

Huang K D, Ye L, Chen L, Wang Q, Dai L, Zhou J, Singh V P, Huang M, Zhang J (2018). Risk analysis of flood control reservoir operation considering multiple uncertainties. J Hydrol (Amst), 565: 672–684

Hu C H, Guo S L, Xiong L H, Peng D (2005). A modified Xinanjiang model and its application in Northern China. Hydrol Res, 36(2): 175–192

Jiang S, Ren L, Hong Y, Yang X, Ma M, Zhang Y, Yuan F (2014). Improvement of multi-satellite real-time precipitation products for ensemble streamflow simulation in a middle latitude basin in south China. Water Resour Manage, 28(8): 2259–2278

Jie M X, Chen H, Xu C Y, Zeng Q, Chen J, Kim J S, Guo S, Guo F Q (2018). Transferability of conceptual hydrological models across temporal resolutions: approach and application. Water Resour Manage, 32(4): 1367–1381

Kang L, Zhou L, Zhang S (2017). Parameter estimation of two improved nonlinear Muskingum models considering the lateral flow using a hybrid Algorithm. Water Resour Manage, 31(14): 4449–4467

Khan M M, Shamseldin A Y, Melville B W, Shoaib M (2015). Stratification of NWP forecasts for medium-range ensemble streamflow forecasting. J Hydrol Eng, 20(7): 04014076

Laiolo P, Gabellani S, Rebora N, Rudari R, Ferraris L, Ratto S, Stevenin H, Cauduro M (2014). Validation of the flood-proofs probabilistic forecasting system. Hydrol Processes, 28(9): 3466–3481

Lerch S, Thorarinsdottir T L (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. Tellus, 65 (10): 98–110

Lewis D, Singer M J, Dahlgren R A, Tate K W (2000). Hydrology in a California oak woodland watershed: a 17-year study. J Hydrol (Amst), 240(1–2): 106–117

Lan T, Lin K, Liu Z, He Y H, Xu C Y, Zhang H B, Chen X H (2018). A clustering preprocessing framework for the subannual calibration of a hydrological model considering climate-land surface variations. Water Resour Res, 54

Lin K, Lv F, Chen L, Singh V P, Zhang Q, Chen X (2014). Xinanjiang model combined with Curve Number to simulate the effect of land use change on environmental flow. J Hydrol (Amst), 519: 3142–3152

Liu J, Xie Z (2014). BMA probabilistic quantitative precipitation forecasting over the Huaihe Basin using TIGGE multi-model ensemble forecasts. Mon Weather Rev, 142(4): 1542–1555

Liu Z, Guo S, Zhang H, Liu D, Yang G (2016). Comparative study of three updating procedures for real-time flood forecasting. Water Resour Manage, 30(7): 2111–2126

Liu Z, Guo S, Xiong L, Xu C Y (2018). Hydrologic uncertainty processor based on copula function. Hydrol Sci J, 63(1): 74–86

Mascaro G, Vivoni E R, Deidda R (2011). Impact of basin scale and initial condition on ensemble streamflow forecast uncertainty. In: The 25th Conference on Hydrology, American Meteorological Society

Najafi M R, Moradkhani H (2016). Towards ensemble combination of seasonal streamflow forecasts. J Hydrol Eng, 21(1): 04015043

Nash J E, Sutcliffe J V (1970). River flow forecasting through conceptual models: part 1: a discussion of principles. J Hydrol (Amst), 10(3): 282–290

Oudin L, Andréassian V, Mathevet T, Perrin C, Michel C (2006). Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resour Res, 42(7): 887–896

Parasuraman K, Elshorbagy A (2007). Cluster-based hydrologic prediction using Genetic Algorithm-trained Neural Networks. J Hydrol Eng, 12(1): 52–62

Perrin C, Michel C, Andréassian V (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. J Hydrol (Amst), 242(3–4): 275–301

Perrin C, Michel C, Andréassian V (2003). Improvement of a parsimonious model for streamflow simulation. J Hydrol (Amst), 279(1–4): 275–289

Raftery A E, Gneiting T, Balabdaoui F, Polakowski M (2005). Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev, 133(5): 1155–1174

Steinschneider S, Brown C (2011). Influences of North Atlantic climate variability on low-flows in the Connecticut River Basin. J Hydrol

(Amst), 409(1–2): 212–224

Sloughter J M, Raftery A E, Gneiting T, Fraley C (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon Weather Rev, 135(9): 3209–3220

Thorarinsdottir T L, Gneiting T (2010). Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. J R Stat Soc (Ser A), 173(2): 371–388

Tian Y, Xu Y P, Zhang X J (2013). Assessment of climate change impacts on river high flows through comparative use of GR4J, HBV and Xinanjiang models. Water Resour Manage, 27(8): 2871–2888

Todini E (2017). Flood forecasting and decision making in the new millennium: where are we? Water Resour Manage, 31(8): 1–19

Wilks D S, Hamill T M (2007). Comparison of ensemble-MOS methods using GFS reforecasts. Mon Weather Rev, 135(6): 2379–2390

WMO (2005) First Workshop on the THORPEX Interactive Grand Global Ensemble (TIGGE), Final Report

WMO (2010) Workshop on the Strategy and Action Plan of the WMO Flood Forecasting Initiative, Final Report

Wu C L, Chau K W (2006). A flood forecasting neural network model with genetic algorithm. Int J Environ Pollut, 28(3–4): 261–273

Wu Z, Wu J, Lu G (2016). A one-way coupled atmospheric-hydrological modeling system with combination of high-resolution and ensemble precipitation forecasting. Front Earth Sci, 10(3): 432–443

Xiong F, Guo S, Chen L, Yin J, Liu P (2018). Flood frequency analysis using Halphen distribution and maximum entropy. J Hydrol Eng, 23 (5): 04018012

Yue S, Ouarda T B M J, Bobée B (2001). A review of bivariate gamma distributions for hydrological application. J Hydrol (Amst), 246(1–4): 1–18

Zhao L, Qi Dan, Tian F, Wu H, D, J, Wang Z, Li A (2012). Probabilistic flood prediction in the upper Huaihe catchment using TIGGE data. J Meteorol Res, 26(1): 62–71

Zhao R (1992). The Xinanjiang model applied in China. J Hydrol (Amst), 135(1–4): 371–381

Zhong Y, Guo S, Ba H, Xiong F, Chang F J, Lin K (2018a). Evaluation of the BMA probabilistic inflow forecasts using TIGGE numeric precipitation predictions based on artificial neural network. Hydrol Res, 49(5): 1417–1433

Zhong Y, Guo S, Liu Z, Wang Y, Yin J (2018b). Quantifying differences between reservoir inflows and dam site floods using frequency and risk analysis methods. Stoch Environ Res Risk Assess, (6):1–15