A Two-Stage Framework for Bias and Reliability Tests of Ensemble Hydroclimatic Forecasts

Tongtiegang Zhao, Shaotang Xiong, Jiahiao Wang, Zhiyong Liu, Denghua Yan, Yongyong Zhang, Xiaohong Chen, and Hao Wang

1 Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Civil Engineering, Sun Yat-sen University, Guangzhou, China; 3 State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing, China; 5 Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

Abstract

The popular probability integral transform (PIT) uniform plot presents informative empirical illustrations of five types of ensemble forecasts, that is, reliable, under-confident, over-confident, negatively biased and positively biased. This paper has built a novel two-stage framework upon the PIT uniform plot to quantitatively examine the forecast attributes of bias and reliability. The first stage utilizes the test statistic on bias to examine whether the mean of PIT values is equal to the theoretical mean of standard uniform distribution. Then, the second stage uses the test statistic on reliability to examine whether the mean squared deviation from the theoretical mean is equal to the theoretical variance of standard uniform distribution. Therefore, by using the two-tailed bootstrap hypothesis testing, the first stage identifies unbiased ensemble forecasts, negatively biased forecasts and positively biased forecasts; the second stage focuses on unbiased ensemble forecasts to furthermore identify reliable forecasts, under-confident forecasts and over-confident forecasts. Numerical experiments are devised for the National Centers for Environmental Prediction’s Climate Forecast System version 2 ensemble forecasts of global precipitation. The results highlight the existence of various shapes of the PIT uniform plots. Due to extreme values of observed precipitation, the PIT uniform plots in some cases can substantially deviate from the 1:1 line even though the mean and variance of ensemble forecasts are respectively in accordance with the mean and variance of observations. Nevertheless, the two-stage framework along with the two test statistics serves as a robust tool for the verification of ensemble hydroclimatic forecasts.

Plain Language Summary

Ensemble hydroclimatic forecasts provide useful information for water resources planning and management. The probability integral transform (PIT) uniform plot presents informative visualizations of five types of ensemble forecasts, that is, reliable, under-confident, over-confident, negatively biased and positively biased. While the use of PIT uniform plot is for individual grid cells, this paper proposes a novel two-stage framework along with two test statistics for the examinations of forecast bias and reliability at the global scale. Specifically, the test of bias in the first stage identifies unbiased, negatively biased and positively biased ensemble forecasts; and then, the test of reliability in the second stage is applied to unbiased ensemble forecasts so as to identify reliable, under-confident and over-confident ensemble forecasts. Numerical experiments are devised to investigate the bias and reliability for ensemble global precipitation forecasts generated by the National Centers for Environmental Prediction’s Climate Forecast System version 2. The five types are effectively identified for raw ensemble forecasts and also for post-processed forecasts generated by the linear scaling, quantile mapping and Gamma-Gaussian models. The results demonstrate the effectiveness of the two-stage framework and highlight that forecast post-processing can play an important part to provide more useful information for decision-making.

1. Introduction

Ensemble hydroclimatic forecasts provide valuable information for water resources planning and management (Cloke & Pappenberger, 2009; Duan et al., 2020; Kopпа et al., 2019; Palash et al., 2020; Turner et al., 2017). Compared to conventional deterministic forecasts that provide the “best” point estimate and therefore disregard forecast uncertainty, ensemble forecasts consist of an ensemble of scenarios, that is, ensemble members, to numerically formulate the distribution of hydroclimatic variables and to explicitly characterize the forecast
Various models have been developed for ensemble hydroclimatic forecasting (Becker et al., 2022; Cloke & Pappenberger, 2009; Hao et al., 2017; Palash et al., 2020; W. Wu et al., 2020). Dynamical models account for the physical hydroclimatic processes in the forecasting (Bennett et al., 2016; Gebrechorkos et al., 2022; Yuan et al., 2018). For example, global climate models (GCMs) couple atmosphere, land, ocean and sea ice modules and assimilate observational information to generate ensemble climate forecasts of global precipitation and temperature (Becker et al., 2022; Johnson et al., 2019; Saha et al., 2014) and furthermore climate forecasts are applied to drive hydrological models to generate ensemble hydrological forecasts of streamflow and soil moisture (Palash et al., 2020; Pechlivanidis et al., 2020; Q. J. Wang et al., 2020). Statistical models identify key explanatory variables of target hydroclimatic variables and quantify the forecast uncertainty to generate ensemble forecasts (Schepen et al., 2020; Q. J. Wang et al., 2009; Zarekarizi et al., 2021). Machine learning models are efficient in dealing with nonlinear dependency relationships and produce hydroclimatic forecasts from massive global and regional datasets (J. Liu et al., 2022; Pan et al., 2022; S. Jiang et al., 2020).

Diagnostic plots have been developed to verify the predictive performance of ensemble hydroclimatic forecasts on a case-by-case basis (Cloke & Pappenberger, 2009; Huang & Zhao, 2022; Keller & Hense, 2011; Murphy, 1993; Wilks, 2011). For ensemble forecasts and the corresponding observations in a given case study, the time series plot presents forecasts and observations against the time index and it can effectively show the predictive performance in periods of extremely high and low events (Bennett et al., 2016; Schepen et al., 2018; Q. J. Wang et al., 2009). The quantile range plot shows ensemble forecasts and the corresponding observations against the ensemble median and it pays attention to the correspondence of high or low values of forecasts to high or low values of observations (Huang et al., 2021; Q. J. Wang et al., 2009; Zhao et al., 2019). Moreover, the probability integral transform (PIT) uniform plot illustrates the capability of ensemble spread in capturing the distribution of observations by empirical visualization of the distribution of PIT values (J. Xu et al., 2022; Koutsoyiannis & Montanari, 2022; Shao et al., 2021; Q. J. Wang et al., 2020; Zhao et al., 2017).

Bias and reliability are two of the most important attributes for the verification of ensemble hydroclimatic forecasts (Huang & Zhao, 2022; Murphy, 1993; Wilks, 2011) and both of them are important determinants of the predictive performance of ensemble forecasts (DelSole & Tippett, 2014; Gilleland et al., 2018; Keller & Hense, 2011). In theory, ensemble forecasts are favorable when they provide reliable information on future hydroclimatic conditions. However, real-world ensemble forecasts can be positively or negatively biased, that is, overall higher or lower than observations. In addition, unbiased ensemble forecasts are not necessarily reliable because they can be over-confident or under-confident if their ensemble spreads are too narrow or too wide to capture the corresponding observations. The existence of different types of ensemble forecasts complicates the process of forecast verification and hinders applications of ensemble hydroclimatic forecasts. Therefore, in this paper, we are motivated to develop a two-stage framework along with two test statistics for the bias and reliability tests of ensemble hydroclimatic forecasts. As will be demonstrated through the methods and results, the two-stage framework effectively extends the use of the popular PIT uniform plot from individual case studies to large-scale applications. The attributes of bias and reliability of raw, bias-corrected and calibrated ensemble GCM precipitation forecasts are effectively illustrated.

2. PIT Uniform Plot of Bias and Reliability

2.1. Formulation of PIT Values

The PIT uniform plot is one of the most popular diagnostic tools to examine the bias and reliability of ensemble hydroclimatic forecasts (Bennett et al., 2016; Koutsoyiannis & Montanari, 2022; Q. J. Wang et al., 2020; Schepen...
Specifically, the PIT uniform plot takes advantage of the multiple members of ensemble forecasts:

\[ F_k = \begin{bmatrix} f_{k,1} & f_{k,2} & \ldots & f_{k,N} \end{bmatrix} \quad (k = 1, 2, \ldots, K) \]  

(1)

in which \( F_k \) is the \( k \)th ensemble forecasts and \( f_{k,n} (n = 1, 2, \ldots, N) \) is the \( n \)th member of \( F_k \). As there are in total \( N \) ensemble members, a cumulative distribution function (CDF) can be formulated either by using the empirical CDF or by explicitly fitting the marginal distribution (Huang et al., 2021; Shao et al., 2021; Zhao et al., 2019). For the observation \( o_k \) corresponding to \( F_k \), the PIT value is derived:

\[ PIT_k = CDF_k (o_k) \]  

(2)

in which \( PIT_k \) represents the value of cumulative probability of \( o_k \) in the CDF(●) of the \( k \)th ensemble forecasts. It is noted that some hydroclimatic variables are censored, for example, precipitation and streamflow are with the lower bound of zero. When \( o_k \) reaches the censoring threshold \( o_{\text{censoring}} \), that is, \( o_k = o_{\text{censoring}} \), the cumulative probability corresponding to the threshold, that is, \( CDF_k (o_{\text{censoring}}) \), is obtained. A pseudo-PIT value is randomly drawn from the uniform distribution bounded by 0 and \( CDF_k (o_{\text{censoring}}) \):

\[ \text{pseudo} - PIT_k \sim U (0, CDF_k (o_{\text{censoring}})) \]  

(3)

While \( CDF_k (o_{\text{censoring}}) \) is a fixed value, \( \text{pseudo} - PIT_k \) in Equation 3 takes the randomness of the PIT value into account for the censored observation (Q. J. Wang et al., 2009; Schepen et al., 2018; Wilks, 2011). If the ensemble forecasts \( F_k \) perfectly capture the observation \( o_k \), then \( o_k \) would statistically be considered as a sample randomly drawn from the ensemble members:

\[ f_{k,1} \quad f_{k,2} \quad \ldots \quad f_{k,N} \]  

(Murphy, 1993; Q. J. Wang et al., 2009; Wilks, 2011). As a result, \( PIT_k \) is expected to follow the standard uniform distribution:

\[ PIT_k \sim U (0, 1) \]  

(4)

With respect to Equation 4, it is pointed out that with only one pair of \( F_k \) and \( o_k \), it is not possible to test the distribution of \( PIT_k \). To bridge the gap, the PIT uniform plot pays attention to \( PIT_k \) \((k = 1, 2, \ldots, K)\) across a series of paired ensemble forecasts \( F_k \) \((k = 1, 2, \ldots, K)\) and observations \( o_k \) \((k = 1, 2, \ldots, K)\). Specifically, the PIT values, as well as pseudo-PIT values, are sorted in ascending order and then plotted against standard uniform variates:

\[ PIT_k^* \sim u_k \quad \left( u_k = \frac{k}{K+1}; k = 1, 2, \ldots, K \right) \]  

(5)

in which \( PIT_k^* \) is the \( k \)th sorted PIT value. The variate \( u_k \) takes the form of the Weibull plotting positions of \( K \) equal-distant numbers located between 0 and 1 to represent the standard uniform variate (Makkonen, 2006). Following the PIT uniform plot, the above formulations of \( PIT_k^* \) \((k = 1, 2, \ldots, K)\) are generally based on the assumption that the PIT values, which are derived from \( K \) pairs of ensemble forecasts and observations, are independent from one another (Murphy, 1993; Q. J. Wang et al., 2009; Schepen et al., 2018; Shao et al., 2021; Wilks, 2011). It is possible that the instances of forecasts and observations exhibit spatial and temporal dependence, as well as contemporaneous dependence among each other, that can make the PIT values not independent and in those cases, more sophisticated methods are in demand to account for the effect of dependence (DelSole & Tippett, 2014; Gilleland et al., 2018).

The PIT uniform plot would be along the 1:1 line if the ensemble forecasts provide perfectly reliable information of the observations. As a result, the difference between \( PIT_k^* \) and \( u_k \) \((k = 1, 2, \ldots, K)\) tells how unreliable the ensemble forecasts are (Renard et al., 2010):

\[ PIT_a = 1 - \frac{2}{K} \sum_{k=1}^{K} \left| PIT_k^* - u_k \right| \]  

(6)

in which \( PIT_a \) is the alpha index measuring the deviation of the PIT uniform plot from the theoretical 1:1 relationship. As can be seen, \( PIT_a \) takes the value of 1 when the sorted PIT values perfectly match the theoretical Weibull plotting positions. The value of \( PIT_a \) gradually decreases as ensemble forecasts become less reliable.
Besides the CDF of PIT values, it is noted that the statistical uniformity can also be investigated by the rank histogram of PIT values (Hamill, 2001). Keller and Hense (2011) proposed to formulate the rank histogram by using the Beta distribution and developed $\alpha$-$\beta$ to characterize the ensemble spread and $\alpha$-$\beta$-$\nu$ to illustrate the ensemble bias.

2.2. Five Types of Ensemble Forecasts

The use of the PIT uniform plots to empirically showcase the bias and reliability of ensemble hydroclimatic forecasts is presented in Figure 1. Conceptually, the five different shapes of PIT uniform plots correspond to overall five types of ensemble forecasts (Koutsoyiannis & Montanari, 2022; Q. J. Wang et al., 2009; Renard et al., 2010):

(a) The first type is reliable ensemble forecasts. In this case, the PIT values as a whole follow the standard uniform distribution:

$$[PIT_1, PIT_2, ..., PIT_K] \sim U(0, 1)$$  \hspace{1cm} (7)

Accordingly, as shown by the green line in Figure 1a, the plot of sorted PIT values against standard uniform variates is along the 1:1 line.

(b) The second type is under-confident ensemble forecasts. The under-confidence means that the ensemble spreads are too wide and that the observations tend to fall in the middle part of the ensemble spreads. In this case, the PIT values are around 0.5:

$$[PIT_1, PIT_2, ..., PIT_K] \rightarrow 0.5$$ \hspace{1cm} (8)

As shown by the cyan line in Figure 1b, the PIT uniform plot for under-confident ensemble forecasts exhibits an inverse-S shape.
(c) The third type is over-confident ensemble forecasts. The over-confidence suggests that the ensemble spreads are too narrow to capture the observations. In this case, the observations frequently fall toward the two tails or out of the ensemble spreads. The PIT values become either 0.0 or 1.0 when the observations fall below (above) the ensemble spread:

\[
\begin{align*}
\{PIT_1, PIT_2, ..., PIT_K\} & \rightarrow \\
& \begin{cases} 
0.0 \\
1.0
\end{cases}
\end{align*}
\]

As shown by the purple line in Figure 1c, the PIT uniform plot for under-confident ensemble forecasts is S-shaped.

(d) The fourth type is negatively biased ensemble forecasts. The negative bias means that ensemble forecasts are overall smaller than observations. In this case, the observations tend to fall in the upper part of ensemble spreads or above the maximum ensemble member. Accordingly, the PIT values tend toward 1.0:

\[
\begin{align*}
\{PIT_1, PIT_2, ..., PIT_K\} & \rightarrow 1.0
\end{align*}
\]

As shown by the brown line in Figure 1d, the PIT uniform plot for negatively biased ensemble forecasts is above the 1:1 line.

(e) The fifth type is positively biased ensemble forecasts. The positive bias indicates that ensemble forecasts are generally larger than observations. Accordingly, the observations tend to fall in the lower part of the ensemble spreads or below the minimum ensemble member. As a result, the PIT values tend toward 0.0:

\[
\begin{align*}
\{PIT_1, PIT_2, ..., PIT_K\} & \rightarrow 0.0
\end{align*}
\]

As shown by the yellow line in Figure 1e, the PIT uniform plot for positively biased ensemble forecasts is below the 1:1 line.

In addition to the conceptual illustrations in Figure 1, five numerical experiments that synthesize observations and ensemble forecasts from pre-specified Gaussian distributions are devised to illustrate how the forecast attributes of bias and reliability determine the shapes of the PIT uniform plots. For more details, please refer to Text S1 in Supporting Information S1. By fixing the mean and variance of observations, the adjustments of the mean and variance of ensemble forecasts facilitate the generation of reliable, under-confident, over-confident, negatively biased and positively biased ensemble forecasts. Overall, the shapes of PIT uniform plots under the five types of ensemble forecasts in Figures S1 to S5 in Supporting Information S1 are observed to follow the conceptual illustrations in Figure 1.

3. Two-Stage Framework for Bias and Reliability Tests

3.1. Two-Stage Framework

In this paper, the proposed two-stage framework focuses on the forecast attributes of bias and reliability that determine the different shapes of the PIT uniform plots (Figure 1). From the perspective of mathematical logic, unbiasedness is a necessary condition of reliability whereas reliability is a sufficient condition of unbiasedness. That is, reliable ensemble forecasts must be unbiased; on the other hand, unbiased ensemble forecasts are not necessarily reliable (they can be under- or over-confident). Furthermore, besides unbiased forecasts, negatively or positively biased ensemble forecasts can be under- or over-confident. For example, negatively or positively biased deterministic forecasts are naturally over-confident (Bennett et al., 2016; Cloke & Pappenberger, 2009; Medina & Tian, 2020). Therefore, as to the five types of ensemble forecasts illustrated by the PIT uniform plot, the first three types, that is, reliable, under-confident and over-confident, are for unbiased forecasts. The last two types, that is, negatively biased and positively biased, are for biased forecasts, of which the characteristics of under- and over-confidence play a secondary role in determining the shapes of the PIT uniform plots.

Overall, the condition of reliability is more stringent than the condition of unbiasedness. By following the principle of simple-to-complex, the two-stage framework in Figure 2 tests the bias in the first stage and the reliability in the second stage:

The first stage pays attention to the forecast attribute of bias. The significance test of bias generally has two outcomes, that is, non-existence of significant bias and existence of significant bias. Furthermore, under the case
of significant bias, the two-tailed test of bias facilitates two sub-outcomes, that is, negative bias and positive bias. It is noted that negatively and positively biased forecasts are not subject to the test of reliability in the second stage and that these two types of forecasts can be under- or over-confident.

The second stage is concentrated on the attribute of reliability for ensemble forecasts that are tested to be unbiased in the first stage. The significance test of reliability also has two outcomes, that is, non-existence of defect in reliability and existence of defect in reliability. In addition, the two-tailed test of reliability yields two sub-outcomes under the existence of defect in reliability, that is, under-confidence and over-confidence.

### 3.2. Tests of Bias and Reliability

The two-stage framework in Figure 2 facilitates an effective procedure to identify the five types of ensemble forecasts illustrated by the PIT uniform plot. As shown by the diamond boxes in the flowchart, the test statistics of bias and reliability are essential for the implementation of the two-stage framework. For the purpose of hypothesis testing, three observations are made for the PIT values that as a whole determine the shapes of the PIT uniform plots in Figure 1. The first observation is that the mean and variance of PIT values for reliable ensemble forecasts are respectively equal to the theoretical mean and variance, which are respectively 0.5 and 0.083, of the standard uniform distribution. The second observation is that the mean of PIT values generally falls above or below 0.5 for negatively or positively biased ensemble forecasts. The third observation is that for under- and over-confident ensemble forecasts, although the mean values of their PIT values are around 0.5, the variance of under- or over-confident forecasts tends to be smaller or larger than the variance of the standard uniform distribution.

Therefore, by selecting 0.5 to be the benchmark, the test statistic on bias, which pays attention to the deviation of PIT values from 0.5, is formulated:

\[
d_m = \frac{1}{K} \sum_{k=1}^{K} (PIT_k - 0.5)
\]  

Figure 2. Flowchart of the two-stage framework along with the tests of bias and reliability to identify the five types of ensemble forecasts.
Furthermore, the test statistic on reliability, which focuses on the squared deviation of PIT values from 0.5, is formulated:

\[ d_v = \frac{1}{K} \sum_{k=1}^{K} (PIT_k - 0.5)^2 \quad (13) \]

The two test statistics in Equations 12 and 13 share the same null hypothesis that the ensemble forecasts \( F_k \) (\( k = 1, 2, \ldots, K \)) perfectly capture the observation \( a_i \) (\( k = 1, 2, \ldots, K \)). It is noted that \( d_m \) and \( d_v \) are generally non-parametric because they are not assumed to come from prescribed models with parameters.

Under the null hypothesis, the observation in theory can be considered as a random sample drawn from the \( N \) members of ensemble forecasts. That is, \( a_i \in_R [f_{k,1} f_{k,2} \ldots f_{k,N}] \) (\( k = 1, 2, \ldots, K \)) in which \( \epsilon_R \) represents the mathematical operation of random selection of a sample from a given set. In this case, the corresponding PIT value takes a random value from the \( N \) Weibull's plotting positions:

\[ PIT_i \in_R \left[ \frac{1}{N+1} \frac{2}{N+1} \ldots \frac{N}{N+1} \right] \quad (k = 1, 2, \ldots, K) \quad (14) \]

in which \( \frac{i}{N+1} \) (\( i = 1, 2, \ldots, N \)) represents the position of the \( i \)th across the \( N \) members of the ensemble forecasts.

The null hypothesis is tested by bootstrapping (Efron, 1992). Specifically, one round of random selection of \( K \) PIT values (Equation 13) facilitates one realization of \( d_m \) and \( d_v \) (Equations 11 and 12); ten thousand rounds facilitate ten thousand realizations of \( d_m \) and \( d_v \). Furthermore, the ten thousand realizations of \( d_m \) and \( d_v \) facilitate the derivations of respective reference distributions of \( d_m \) and \( d_v \):

\[ \begin{bmatrix} d_m^1 & d_m^2 & \ldots & d_m^{10000} \\ d_v^1 & d_v^2 & \ldots & d_v^{10000} \end{bmatrix} \Rightarrow CDF_m(d_m) \quad CDF_v(d_v) \quad (15) \]

In Equation 13, \( CDF_m(d_m) \) and \( CDF_v(d_v) \) respectively represent the cumulative distribution function of \( d_m \) and \( d_v \) under the null hypothesis.

### 3.3. Five Types of Ensemble Forecasts

For the ensemble forecasts and observations in a given case study, the two calculated test statistics of bias and reliability are respectively denoted as \( d_m^* \) and \( d_v^* \). For a pre-specified significance level of \( \alpha \), which is set to be 0.05 in the analysis, the two-stage framework is implemented to identify which type the ensemble forecasts are.

In the first stage, \( d_m^* \) is compared to the percentiles obtained from the inverse cumulative distribution function \( CDF_m^{-1}(d_m) \). The case of unbiasedness is indicated by the result that \( d_m^* \) falls into the \( 1 - \alpha \) confidence interval:

\[ CDF_m^{-1}\left(\frac{\alpha}{2}\right) \leq d_m^* \leq CDF_m^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (16) \]

The above condition indicates unbiased ensemble forecasts. If the above condition is not satisfied, then the ensemble forecasts are identified to be biased. There are two cases under the two-tailed test. One case is that \( d_m^* \) is above the \( 1 - \frac{\alpha}{2} \) percentile:

\[ d_m^* > CDF_m^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (17) \]

The above condition indicates negatively biased ensemble forecasts. The other case is that \( d_m^* \) is below the \( \frac{\alpha}{2} \) percentile:

\[ d_m^* < CDF_m^{-1}\left(\frac{\alpha}{2}\right) \quad (18) \]

The above condition indicates positively biased ensemble forecasts.
In the second stage, $d_v^*$ is compared to $CDF_v(d_v)$ for ensemble forecasts that are tested to be unbiased (Equation 16) in the first stage (Figure 2). The null hypothesis of reliability is tested to be true if $d_v^*$ falls into the $1 - \alpha$ confidence interval:

$$CDF_v^{-1}\left(\frac{\alpha}{2}\right) \leq d_v^* \leq CDF_v^{-1}\left(1 - \frac{\alpha}{2}\right)$$ (19)

The above condition indicates reliable ensemble forecasts. If the above condition is not satisfied, then the two-tail test derives another two cases. One case is that $d_v^*$ is below the $\frac{\alpha}{2}$ percentile:

$$d_v^* < CDF_v^{-1}\left(\frac{\alpha}{2}\right)$$ (20)

The above condition indicates under-confident ensemble forecasts. The other case is that $d_v^*$ is above the $1 - \frac{\alpha}{2}$ percentile:

$$d_v^* > CDF_v^{-1}\left(1 - \frac{\alpha}{2}\right)$$ (21)

The above condition indicates over-confident ensemble forecasts.

Therefore, the five types of ensemble forecasts are quantitatively identified through the bootstrap hypothesis testing of $d_v^*$ and $d_m^*$. To summarize, Equations 16 and 19 are the conditions of reliable ensemble forecasts; Equations 16 and 20 the conditions of under-confident ensemble forecasts; Equations 16 and 21 the conditions of over-confident ensemble forecasts; Equation 17 the condition of negatively biased ensemble forecasts; and Equation 18 the condition of positively biased ensemble forecasts.

4. Numerical Experiments

4.1. Experiments of Raw Forecasts

The two-stage framework along with the two test statistics are utilized to examine the bias and reliability for the raw ensemble precipitation forecasts generated by the National Centers for Environmental Prediction (NCEP)'s Climate Forecast System version 2 (CFSv2) (Saha et al., 2014). The global precipitation forecasts are generally five-dimensional:

$$F = [f_{s,l,n,y,x}]$$ (22)

in which $F$ represents the data set of NCEP-CFSv2 forecasts and $f$ individual forecast value. The forecasts, which are at the monthly timescale, are specified by five dimensions: (a) start time $s$, which is the month at which forecasts are produced, for example, January 1982, February 1982, March 1982 and so on; (b) lead time $l$, which is the number of months ahead the start time and ranges from 0 to 9, that is, 10 lead times in total; (c) ensemble member $n$, which ranges from 1 to 24 and presents a 24-member numerical characterization of forecast uncertainty; (d) latitude $y$, which ranges from $-90$ to $90$; and (e) longitude $x$, which ranges from 0 to 359.

The observed precipitation corresponding to the forecasts is sourced from the Climate Prediction Center (CPC)'s Unified Rain-gauge Database (URD) (Xie et al., 2007). It is three-dimensional:

$$O = [o_{t,y,x}] \quad (t = s + l)$$ (23)

in which $O$ is the data set of CPC-URD observations and $o$ individual observation value. The three dimensions are (a) target time $t$, which adds lead time $l$ to start time $s$ in aligning forecasts with observations; (b) latitude $y$; and (c) longitude $x$.

The NCEP-CFSv2 precipitation forecasts targeting at June-July-August (JJA), which is generally summer in the Northern Hemisphere and winter in the Southern Hemisphere, are investigated in the experiment. The start time is selected to be June. That is, the raw forecasts, which are generated in June 1982, June 1983, … and June 2010,
are pooled and the monthly forecasts at the lead times of 0, 1 and 2 months are aggregated into seasonal forecasts of JJA precipitation:

\[ F_{\text{JJA},\text{month}} = [f_{k,n}] \quad (k = 1982, 1983, ..., 2010; n = 1, 2, ..., 24) \]  \hspace{1cm} (24)

in which \( k \) is the year in the CFSv2 hindcast period from 1982 to 2010 (Becker et al., 2022; Saha et al., 2014) and \( n \) is the index of ensemble member. In accordance with the forecasts, the monthly observations are also aggregated into seasonal:

\[ O_{\text{JJA},\text{month}} = [o_k] \quad (k = 1982, 1983, ..., 2010) \]  \hspace{1cm} (25)

Therefore, \( \text{PIT}_k \) is calculated by comparing \( o_k \) to \([f_{k,1}, f_{k,2}, ..., f_{k,24}]\) and then \( d^o_n \) and \( d^* \) are obtained:

\[ \text{PIT}_k \quad (k = 1982, 1983, ..., 2010) \Rightarrow \begin{cases} d^o_n \\ d^* \end{cases} \hspace{1cm} (26) \]

For a selected grid cell \((y, x)\), the two-stage framework along with the two test statistics (Figure 2) tells the type of the ensemble forecasts based on \( d^o_n \) and \( d^* \) (Equations 16–21). Applying the above tests one by one to grid cells around the world, the bias and reliability are examined for the ensemble NCEP-CFSv2 forecasts of global precipitation. It is noted that the tests are applicable to PIT (Equation 2) and pseudo-PIT values (Equation 3) and that grid cells with zero values of precipitation are not considered in the experiment for the sake of simplicity.

### 4.2. Experiments of Forecast Post-Processing

Raw GCM forecasts are known to exhibit biases and have defects in reliability (Bauer et al., 2015; Becker et al., 2022; Schepen et al., 2020; Shao et al., 2021; Zhao et al., 2017). Therefore, besides the examinations of bias and reliability for raw forecasts, this paper furthermore employs three post-processing models, that is, the linear scaling, the quantile mapping and the Gamma-Gaussian model, to improve the raw forecasts. The forecast attributes of bias and reliability are re-examined and then compared to those of raw forecasts.

The linear scaling focuses on the difference between the mean value of forecasts and the mean value of observations. It applies a simple scaling coefficient \( s \) to correct the bias in raw forecasts (Crochemore et al., 2016; Teutschbein & Seibert, 2013):

\[ f_{k,n}^{ls} = s \times f_{k,n} \quad (n = 1, 2, ..., 24) \]  \hspace{1cm} (27)

in which \( f_{k,n}^{ls} \) and \( f_{k,n} \) are respectively linearly scaled and raw forecasts. The coefficient \( s \) is calculated as the ratio of the sum of observations over the sum of forecasts:

\[ s = \frac{\sum_{i=1}^{K} o_i}{\sum_{i=1}^{K} \sum_{k=1}^{N} f_{k,i}} \hspace{1cm} (28) \]

As can be seen, the coefficient \( s \) is larger than 1 when forecasts are overall smaller than observations (Equation 28); as a result, the raw forecast is enlarged (Equation 27). On the other hand, the coefficient is smaller than 1 when forecasts are overall larger than observations and the raw forecast is reduced. In the experiment, the linear scaling is implemented by using the leave one out cross validation (LOOCV). That is, the observation and forecasts in the year \( k \) under investigation are excluded from the calculation of the coefficient \( s \), that is, \( i \neq k \) in Equation 28. In this way, the samples for model fitting and testing are separated and the effect of linear scaling is not artificially inflated (Murphy, 1993; Wilks, 2011). While precipitation exhibits a lower bound of zero, the linear scaling retains this lower bound and adjust the range of raw forecasts to make their mean match up to the mean of observations.
The quantile mapping pays attention to the marginal distributions of forecasts and observations (Gudmundsson et al., 2012; Maraun, 2013; Piani et al., 2010):

\[ f_{k,n}^m = CDF^{-1}_O (CDF_T (f_{k,n})) \]  

(29)

in which \( f_{k,n}^m \) and \( f_{k,n} \) are respectively quantile-mapped and raw forecasts. \( CDF_T (\bullet) \) is the CDF of all ensemble members of raw forecasts and \( CDF^{-1}_O (\bullet) \) is the inverse CDF of observations. Applying the CDF of observations, that is, \( CDF_O (\bullet) \), to both sides of Equation 28, it can be derived that:

\[ CDF_O \left( f_{k,n}^m \right) = CDF_T (f_{k,n}) \]  

(30)

As can be seen, the core idea of quantile mapping is to make the percentile of quantile-mapped forecast in \( CDF_O (\bullet) \) the same as the percentile of raw forecast in \( CDF_T (\bullet) \). In this way, the quantile mapping modifies the marginal distribution of raw forecasts (Gudmundsson et al., 2012; Maraun, 2013; Piani et al., 2010). In the analysis, the CDFs are fitted by the Gamma distribution. Similar to the linear scaling, the quantile mapping is performed by using the LOOCV. That is, the observation and ensemble forecasts in year \( k \) are excluded when fitting \( CDF_T (\bullet) \) and \( CDF^{-1}_O (\bullet) \) to improve ensemble forecasts in year \( k \).

The Gamma-Gaussian model utilizes the Gamma distribution to fit the marginal distributions of ensemble means and observations and then employs the bi-variate Gaussian distribution to characterize their dependency relationship (Huang et al., 2021, 2022). The Gamma distribution facilitates the data normalization:

\[
\begin{align*}
\tilde{f} & = CDF^-1_{\text{Normal}} \left( CDF_T (f) \right) \\
\tilde{o} & = CDF^-1_{\text{Normal}} (CDF_O (o))
\end{align*}
\]  

(31)

As can be seen, the ensemble mean \( \tilde{f} \) (Rougier, 2016) is transformed to standard uniform variate by its CDF and then to standard normal variate by the inverse CDF of standard normal distribution, that is, \( f \sim N (0, 1^2) \).

Similarly, the observation \( o \) is normalized by its CDF and the inverse CDF of standard normal distribution, that is, \( o \sim N (0, 1^2) \). After the normalization, the bi-variate Gaussian distribution is applied to formulate the joint distribution of normalized ensemble mean and observation:

\[
\begin{bmatrix}
\tilde{f} \\
\tilde{o}
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)
\]  

(32)

From the joint distribution, the conditional distribution of normalized observation \( \tilde{o}_k \) upon a given normalized ensemble mean \( \tilde{f}_k \) is derived:

\[ \tilde{o}_k \sim N \left( \rho \tilde{f}_k, 1 - \rho^2 \right) \]  

(33)

As is illustrated, the influence of \( \tilde{f}_k \) on the conditional distribution is determined by how well it correlates with \( \tilde{o} \).

The increase in \( \rho \) not only makes \( \tilde{f}_k \) more effective in determining the conditional mean but also leads to a reduction in the conditional variance. The conditional distribution of observation inferred from the ensemble mean is applied to produce more skillful ensemble forecasts from raw forecasts (Huang et al., 2021, 2022). It is pointed out that the LOOCV also applies to the Gamma-Gaussian model. That is, when calibrating the raw forecasts in year \( k \) (Equation 33), the Gamma and bi-variate Gaussian distributions (Equations 31 and 32) are fitted by using the samples of raw forecasts and observations in years other than \( k \).

5. Results

5.1. Five Types of Ensemble Forecasts

The two-stage framework identifies the five types of ensemble precipitation forecasts generated by the NCEP-CFSv2. In Figure 3, the spatial distribution of the five types is illustrated by different colors at the global scale. Overall, it can be observed that while the NCEP-CFSv2 provides informative forecasts of global precipitation, its raw forecasts are beset by considerable bias and unreliable ensemble spread (Becker et al., 2022;
The orange areas show that there are positively biased ensemble forecasts for more than half of grid cells. Specifically, the raw NCEP-CFSv2 forecasts tend to over-estimate observed precipitation for 61.37% of the grid cells. In the meantime, the brown areas indicate that the raw forecasts present under-estimation of observed precipitation for 23.04% of the grid cells. Overall, there are unbiased ensemble forecasts for 15.61% of grid cells around the world. Among them, 10.37% are with reliable ensemble forecasts (green areas), 0.40% with under-confident forecasts (cyan areas) and 4.84% with over-confident forecasts (purple areas). From Figure 3, one implication is that bias correction and forecast calibration would be a necessary step to improve raw forecasts for practical applications of GCM forecasts (Huang et al., 2021; Li et al., 2017; Schepen et al., 2020; Shao et al., 2021; Zhao et al., 2017).

The PIT uniform plots under the five types of ensemble forecasts are pooled and shown in Figure 4. The color scheme of Figure 4 is the same as those of Figures 1 and 3. In Figure 1 are the conceptual PIT uniform plots for the five types of ensemble forecasts. Comparing Figures 4–1, one remarkable finding is that the PIT uniform plots of real-world forecasts exhibit various shapes. The two-stage framework proposed in this paper deals with the variability. In Figure 4a, it can be observed that the PIT uniform plots, which are in the form of line plots, are around the 1:1 line for ensemble forecasts that are tested to be reliable. In Figure 4b, the PIT uniform plots tend to be inverse-S shaped for under-confident ensemble forecasts. By contrast, the PIT uniform plots in Figure 4c tend to be S shaped for over-confident ensemble forecasts. Furthermore, the PIT uniform plots in Figure 4d fall above the 1:1 line for negatively biased ensemble forecasts, whereas the PIT uniform plots in Figure 4e fall below the 1:1 line for positively biased ensemble forecasts. The alpha index for the five types of forecasts is illustrated by boxplots in Figure 4f. It can be seen that reliable ensemble forecasts generally exhibit the highest alpha index, followed by under- and over-confident forecasts and then by positively and negatively biased ensemble forecasts.

5.2. Raw Ensemble Forecasts at Selected Grid Cells

Under each type of ensemble forecasts, diagnostic plots are generated for the grid cell with the highest value of alpha index in Figure 5. The highest alpha index means that the corresponding PIT uniform plot is the most similar to the 1:1 line for ensemble forecasts in relation to observations. There are three columns in Figure 5. The time series plots are presented in the first column, the quantile range plots in the second column and the PIT uniform plots in the third column. It is noted that instead of line plots in Figures 1 and 4, the PIT uniform plots

Figure 3. Spatial distribution of the five types of National Centers for Environmental Prediction Climate Forecast System version 2 ensemble forecasts of global precipitation in June-July-August.
in Figure 5 are in the form of dot plots to highlight the individual ranked PIT values (Huang & Zhao, 2022; Q. J. Wang et al., 2009; Zhao et al., 2019). The five rows respectively showcase the five types of ensemble forecasts:

1. The first row illustrates ensemble forecasts that are identified to be reliable by the two-stage framework. Both the time series plot and the quantile range plot suggest that the mean value of ensemble forecasts is similar to the mean value of observations, implying unbiasedness. In addition, the ensemble spreads tend to capture the observations. As a result, the PIT uniform plot generally follows the 1:1 line.

2. The second row presents under-confident ensemble forecasts. It can be observed that although the mean values are similar for ensemble forecasts and observations, the ensemble spreads tend to be too wide. Specifically, the time series plot suggests that there is only 1 observation falling above the 90% percentile of ensemble spread among the 29 observations and that there is only 1 observation below the 10% percentile. Overall, the PIT values tend to distribute toward 0.5, leading to a slightly inverse-S shaped PIT uniform plot.

3. The third row is for over-confident ensemble forecasts. Both the time series plot and the quantile range plot show that the ensemble spreads are too narrow to capture the observations, though the mean value of ensemble forecasts is similar to that of observations. The overly narrow ensemble spread makes some PIT values close to 0 and 1, resulting in a slightly S-shaped PIT uniform plot.

4. The fourth row shows negatively biased ensemble forecasts. Through the quantile range plot, it can be observed that the observations tend to be above the 1:1 line. The implication is that the observations are generally larger than the corresponding ensemble forecasts. As a result, the PIT values tend toward 1 and the PIT uniform plot is above the 1:1 line.

5. The fifth row presents positively biased ensemble forecasts. The observations tend to be smaller than the ensemble forecasts. In particular, the time series plot shows that 5 of the 29 observations are below the 10% percentile of ensemble forecasts and that none is above the 90% percentile. Accordingly, the PIT values tend toward 0 and the PIT uniform plot is below the 1:1 line.
Figure 5. Diagnostic plots of the five types of ensemble forecasts at grid cells with the highest alpha indices. The first and second columns are respectively the time series and quantile range plots, in which the red dots represent observations and the black marks, deep blue bars and light blue bars respectively represent the median, [25%, 75%] and [10%, 90%] inter-quantile ranges. The third column is the probability integral transform uniform plots, of which the different colors correspond to the five types of ensemble forecasts in Figure 1.
Ensemble forecasts that are identified to be reliable by the two-stage framework are furthermore investigated. In contrast to Figure 5a that pays attention to the grid cell with the highest alpha index, Figure 6 illustrates the reliable ensemble forecasts at the three grid cells with the lowest, the second lowest and the third lowest alpha indices. From the time series and quantile ranges plots, it can be observed though the ensemble forecasts overall capture the observations, there exist some anomalously large observations that are beyond the ensemble spreads. In the meantime, the subplots in the third column of Figure 6 suggest that the PIT uniform plots are partly along the 1:1 line and that some deviations are observed. The test statistics are examined for the three grid cells. It is noted that $[CDF_m^{-1}(0.025), CDF_m^{-1}(0.975)]$ is [-0.100, 0.100] for $d_m$ and that $[CDF_v^{-1}(0.025), CDF_v^{-1}(0.975)]$ is [0.054, 0.103] for $d_v$. As can be seen, either $d_m^*$ or $d_v^*$ tends to be close to the bounds of the confidence intervals at the three grid cells. The implication is that with the adjustment of significance level, ensemble forecasts at the

---

Figure 6. Diagnostic plots of reliable ensemble forecasts at the three grid cells exhibiting the lowest, second lowest and third lowest alpha indices.
three selected grid cells are likely to be classified into other types. Nevertheless, under the pre-specified significance level of 0.05, the ensemble forecasts at the three grid cells are deemed to be reliable. Overall, Figure 6 is in accordance with Figure 4a and points to the remarkable variability of PIT uniform plots of real-world ensemble forecasts.

5.3. Comparison With Direct Tests of Mean and Variance

As a comparison with the proposed tests of bias and reliability, the ensemble forecasts are also tested by directly using the mean and variance that are respectively the first and second moments of a statistical distribution (Pham, 2006; Wilks, 2011). Specifically, as shown in Figure S18 in Supporting Information S1, the comparison of the mean values of ensemble forecasts and observations in the first stage illustrates whether the ensemble forecasts are biased; the comparison of the variances in the second stage tells the reliability of ensemble forecasts; and therefore, the direct tests of mean and variance also facilitate the identifications of the five types of ensemble forecasts. The spatial distribution of the five types is shown in Figure 7 and the corresponding PIT uniform plots are illustrated in Figure 8. Comparing Figures 7–3, it can be observed that the spatial distributions are to a large extent similar. On the other hand, the comparison of Figure 8 with Figure 4 indicates the existence of remarkable differences in the PIT uniform plots. For the type of reliable ensemble forecasts, the PIT uniform plots evidently exhibit more deviations from the 1:1 line in Figure 8a than in Figure 4a. At some grid cells, the PIT uniform plots are even seen to fall wholly below or above the 1:1 line. For the other four types of ensemble forecasts, the PIT uniform plots are also shown to be more variable in Figures 8b–8e than in Figures 4b–4e. Furthermore, from the boxplots in Figures 8f and 4f, it can be observed that the direct tests of mean and variance lead to some outliers that are indicated by small values of alpha index in particular under the type of reliable ensemble forecasts.

For ensemble forecasts identified to be reliable by the direct tests of mean and variance, three grid cells with the lowest, second lowest and third lowest alpha index are selected to examine the corresponding ensemble forecasts and observations. The layout of Figure 9 is the same as that of Figure 6. In Figure 9, some similar patterns are observed for the three grid cells. Specifically, the observed precipitation is extremely large in 1 year or 2 years but is comparatively low in the other years. Therefore, the identification of reliable ensemble forecasts by the direct tests of mean and variance is attributable to the fact that the ensemble forecasts happen to exhibit similar mean and variance to the observations. On the other hand, the time series and quantile range plots in Figure 9 indicate that the similarities in mean and variance are at the cost of ensemble forecasts drifting toward the extremely large observed precipitation in some years but deviating considerably from observed precipitation in other years. As a result, the corresponding PIT uniform plots drop substantially below the 1:1 line, leading to the low values of
alpha index. Overall, the comparison of Figures 9–6 indicates that the direct tests of mean and variance are more subject to the influence of extreme values.

5.4. Bias and Reliability of Post-Processed Forecasts

The forecast post-processing models of linear scaling, quantile mapping and Gamma-Gaussian are utilized to improve the raw CFSv2 ensemble precipitation forecasts. The results of bias and reliability for the post-processed forecasts by the linear scaling are shown in Figure 10. At the left-hand side of the figure is the spatial distribution of the five types of ensemble forecasts identified by the tests of bias and reliability in Figure 2. At the right-hand side of the figure is the chord diagram illustrating how the percentages of the five types transit from raw forecasts (lower part) to post-processed forecasts (upper part). Comparing Figures 3–10, it can be observed that the correction of the mean value by the linear scaling evidently reduces the percentage of negatively biased forecasts from 23.03% to 2.64% and the percentage of positively biased forecasts from 61.37% to 2.37%. Due to the correction of bias, the percentage of reliable forecasts increases from 10.37% to 45.66%. These results highlight the effectiveness of linear scaling in correcting the bias and indicates the usefulness of linear scaling when bias is the main issue of raw forecasts. On the other hand, it is noted that the linear scaling is less effective in improving the reliability. From Figure 10, it is observed that there are over-confident forecasts in 46.92% of grid cells around the world. Moreover, from the diagnostic plots in Figure S22 in Supporting Information S1, it can be seen that the similar mean values for raw forecasts and observations can make the linear scaling fail to improve raw forecasts.

The results of bias and reliability of post-processed ensemble forecasts generated by the quantile mapping are shown in Figure 11. It can be observed that the percentages of negatively and positively biased forecasts respectively become 0.01% and nearly 0.00% and that owing to the correction of bias, the percentage of reliable forecasts increases to 81.49%. While the quantile mapping is more complicated compared to the linear scaling, these results suggest that the quantile mapping is more effective in correcting the bias. The effectiveness is
attributable to the match-up of the marginal distributions, that is, climatology, for forecasts and observations (Crochemore et al., 2016; Gudmundsson et al., 2012; Maraun, 2013).

The results of bias and reliability of post-processed ensemble forecasts produced by the Gamma-Gaussian model are illustrated in Figure 12. The percentage of reliable forecasts increases to 87.86% from 10.37% for raw forecasts. The percentages of over- and under-confident forecasts are respectively 11.95% and 0.18%. The percentages of negatively and positively biased forecasts are respectively 0.01% and 0.00%. Overall, among the three models of forecast post-processing, the Gamma-Gaussian model tends to be the most effective in correcting the bias and improving the reliability. The effectiveness is attributable to that the Gamma distribution is devised to formulate the marginal distribution of precipitation (Crochemore et al., 2016; Maraun, 2013; Piani et al., 2010) and that

---

**Figure 9.** Diagnostic plots of reliable ensemble forecasts identified by the two-stage framework along with the direct tests of mean and variance. The plots are for the three grid cells exhibiting the lowest, second lowest and third lowest alpha indices.
the bi-variate Gaussian distribution is employed to account for how indicative raw forecasts are of observations (Duan et al., 2020; Huang et al., 2021, 2022).

6. Discussion

The PIT uniform plots presented in Figure 1 provide informative visualization of bias and reliability for ensemble hydroclimatic forecasts (J. Xu et al., 2022; Koutsoyiannis & Montanari, 2022; Q. J. Wang et al., 2020; Shao et al., 2021; Zhao et al., 2017). In this paper, the two-stage framework along with two test statistics are developed to explicitly identify the five types of ensemble forecasts categorized by the PIT uniform plots. It effectively extends the use of PIT uniform plot for individual grid cells to applications at the global scale. Numerical experiments are devised for the investigation of bias and reliability for the NCEP-CFSv2 ensemble forecasts of global precipitation (Saha et al., 2014). The results generally highlight the variability of the PIT uniform plots for real-world ensemble forecasts. In Figures 4 and 7, it is observed that the alpha index of reliable ensemble forecasts is not necessarily higher than the alpha indices of under-confident (over-confident) ensemble forecasts and negatively (positively) biased ensemble forecasts. Furthermore, in Figure 9, it is seen that the PIT uniform plots in some cases can substantially deviate from the 1:1 line although the mean and variance of ensemble forecasts are respectively in accordance with the mean and variance of observations. These outcomes are attributable to the existence of extreme values of observed precipitation that substantially influence the results of bias and reliability. Nevertheless, the two test statistics are demonstrated to be robust tools for the examinations of bias and reliability. Their effectiveness is attributable to the fact that the PIT calculation employs the CDF of
ensemble forecasts to normalize the observation and therefore mitigates the influence of extreme values (C. Jiang et al., 2015; Koutsoyiannis & Montanari, 2022; Q. J. Wang et al., 2020).

Two extended experiments are conducted to investigate the effects of the number of events (Figures S6 to S11 in Supporting Information S1) and the number of ensemble members (Figures S12–S17 in Supporting Information S1) on the results of bias and reliability. The experiments are based on synthetic observations and ensemble forecasts drawn from pre-specified Gaussian distributions (Text S1 and Figures S1–S5 in Supporting Information S1). Notably, reliable ensemble forecasts are generated from the climatological distribution of observations and the other four types of forecasts are generated by varying the mean or variance. From Figures S6a to S11a in Supporting Information S1, it can be observed that as the number of events is reduced from 50 to 10, ensemble forecasts randomly generated from the perfectly reliable climatological distribution become more likely to be identified as not reliable. This result implies the type I error, that is, rejecting the null hypothesis when it is actually true and highlights the importance of the number of events in reducing the type I error. In the meantime, the type II error, that is, failing to reject the null hypothesis when it is actually false, is also observed (Figure S6–S11 in Supporting Information S1). In Figures S12–S17 in Supporting Information S1, the effect of the number of ensemble members on the results of bias and reliability is illustrated. As the number of ensemble members is reduced from 50 to 10, ensemble forecasts generated from the reliable climatological distribution tend to remain reliable, albeit the existence of some type I and type II errors. Overall, the comparison of Figures S17–S11 in Supporting Information S1 suggests that the results of bias and reliability are more sensitive to the number of events than to the number of ensemble members. The indication is that efforts are in demand to augment the events of forecasts and observations to facilitate more effective tests of the forecast attributes of bias and reliability.

Ensemble hydroclimatic forecasts provide valuable information for water resources management (Gebrechorkos et al., 2022; Giuliani et al., 2020; H. Wang et al., 2016; P. Liu et al., 2011; Turner et al., 2017). From the perspective of physically based modeling, it is important to investigate the drivers of the patterns shown in Figures 3–9 to yield insights into the performances of GCM forecasts. The existences of under-/over-confidence and negative/positive biases can relate to atmosphere, land, ocean and sea ice modules and also to data assimilation systems of GCMs (Bauer et al., 2015; Becker et al., 2022; Johnson et al., 2019; Kirtman et al., 2014; Saha et al., 2014). From the perspective of decision-making, ensemble forecasts, which can be generated by dynamical models and also by statistical and machine learning models, are generally taken as ad hoc inputs for various stochastic optimization models (B. W. Xu et al., 2020; Xu et al., 2020; Zhao et al., 2012). The attributes of bias and reliability inevitably affect the management decisions in that positive biases can lead to over-optimistic estimations of water availability and that negative biases can make water availability estimations over-pessimistic. Moreover, under-confidence can cause too conservative water use decisions and over-confidence too bold decisions. As is illustrated by the PIT uniform plots, there exist five types of ensemble forecasts, that is, reliable, under-confident, over-confident, negatively biased and positively biased. While the different types complicate the use of ensemble forecasts,
the two-stage framework facilitates the examinations of bias and reliability of ensemble forecasts in large-scale applications. It can serve as a useful screening and assessment tool for forecast users and water managers before applying ensemble hydroclimatic forecasts to decision-making.

7. Conclusions

This paper has developed a novel two-stage framework along with two test statistics of PIT values for the tests of bias and reliability of ensemble hydroclimatic forecasts. These tools are built upon the PIT uniform plot that has been one of the most popular diagnostic tools of forecast verification. The core idea of this plot is to assess whether the PIT values, which indicate the cumulative probabilities of observations in the CDFs of ensemble forecasts, as a whole follow the standard uniform distribution. The two-stage framework developed in this paper, which is comprised of the test of bias in the first stage and the test of reliability in the second stage, presents a quantitative formulation of the distribution of PIT values in relation to the standard uniform distribution. Specifically, the test of bias examines whether the mean of PIT values is equal to the theoretical mean of standard uniform distribution to identify unbiased ensemble forecasts, negatively biased forecasts and positively biased forecasts; furthermore, the test of reliability examines whether the squared deviation of PIT values from the theoretical mean is equal to the theoretical variance of standard uniform distribution to identify reliable forecasts, under-confident forecasts and over-confident forecasts. Through the numerical experiments of the raw NCEP-CFSv2 ensemble forecasts of global precipitation and three sets of post-processed forecasts generated by the linear scaling, quantile mapping and Gamma-Gaussian models, the usefulness of the two-stage framework is demonstrated and the effectiveness of the two test statistics in comparison to the direct tests of mean and variance is illustrated. In the future, more efforts can be devoted into investigations of real-world ensemble hydroclimatic forecasts to test the robustness of the two-stage framework along with the two test statistics and to facilitate applications of ensemble forecasts to water resources management.

Data Availability Statement

Both the NCEP-CFSv2 ensemble precipitation forecasts and the CPC-URD precipitation observations are downloaded from the North American Multimodel Ensemble (NMME) experiment (https://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/).

References

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 523(7567), 47–55. https://doi.org/10.1038/nature14956

Becker, E. J., Kirtman, B. P., L’Heureux, M., Muñoz, Á. G., & Pegion, K. (2022). A decade of the North American Multimodel Ensemble (NMME): Research, application, and future directions. *Bulletin of the American Meteorological Society*, 103(3), E973–E995. https://doi.org/10.1175/BAMS-D-20-0327.1

Bennett, J. C., Wang, Q. J., Li, M., Robertson, D. E., & Schepen, A. (2016). Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model. *LONG-RANGE ensemble streamflow forecasts. Water Resources Research*, 52(10), 8238–8259. https://doi.org/10.1002/2016WR019193

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375(3–4), 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Crochemore, L., Ramos, M.-H., & Pappenberger, F. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, 20(9), 3601–3618. https://doi.org/10.5194/hess-20-3601-2016

DeSoile, T., & Tippett, M. K. (2014). Comparing forecast skill. *Monthly Weather Review*, 142(12), 4658–4678. https://doi.org/10.1175/MWR-D-14-00045.1

Duan, Q., Pappenberger, F., Wood, A., Cloke, H. L., & Schaake, J. C. (Eds.) (2020). *Handbook of hydrometeorological ensemble forecasting*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40457-3

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 569–593). Springer New York. https://doi.org/10.1007/978-1-4612-4380-9_41

Emerton, R., Zsoter, E., Arnal, L., Cloke, H. L., Marazo, D., Puddicombe, C., ... & (2018). Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-seasonal v1.0. *Geoscientific Model Development*, 11(8), 3327–3346. https://doi.org/10.5194/gmd-11-3327-2018

Gebrechorkos, S. H., Pan, M., Beck, H. E., & Sheffield, J. (2022). Performance of state-of-the-art C3S European seasonal climate forecast models for mean and extreme precipitation over Africa. *Water Resources Research*, 58(3). https://doi.org/10.1029/2021WR031480

Gilleland, E., Hering, A. S., Fowler, T. L., & Brown, B. G. (2018). Testing the tests: What are the impacts of incorrect assumptions when applying confidence intervals or hypothesis tests to compare competing forecasts? *Monthly Weather Review*, 146(6), 1685–1703. https://doi.org/10.1175/MWR-D-17-0295.1

Giuliani, M., Crochemore, L., Pechlivanidou, I., & Castelletti, A. (2020). From skill to value: Isolating the influence of end user behavior on seasonal forecast assessment. *Hydrology and Earth System Sciences*, 24(12), 5891–5902. https://doi.org/10.5194/hess-24-5891-2020

Acknowledgments

We are grateful to the editor, the associate editor and the three anonymous reviewers for the insightful and constructive comments that led to substantial improvements of the quality of the paper. This research is supported by the National Key Research and Development Program of China (2021YFC3001000), the National Natural Science Foundation of China (51979295, 51861125203, 51725905, 52130907 and U1911204) and the Guangdong Provincial Department of Science and Technology (2019ZT08G090).
Wang, H., Lei, X., Guo, X., Jiang, Y., Zhao, T., Wang, X., & Liao, W. (2016). Multi-Reservoir system operation theory and practice. In L. K. Wang, C. T. Yang, & M.-H. S. Wang (Eds.), Advances in water resources management (pp. 1–110). Springer International Publishing. https://doi.org/10.1007/978-3-319-22924-9_1

Wang, Q. J., Bennett, J. C., Robertson, D. E., & Li, M. (2020). A data censoring approach for predictive error modeling of flow in ephemeral rivers. Water Resources Research, 56(1). https://doi.org/10.1029/2019WR026128

Wang, Q. J., Robertson, D. E., & Chiew, F. H. S. (2009). A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites: Seasonal forecasting of streamflows. Water Resources Research, 45(5). https://doi.org/10.1029/2008WR007355

Wilk, D. S. (2011). Forecast verification. In International Geophysics, (Vol. 100, pp. 301–394). Elsevier. https://doi.org/10.1016/B978-0-12-385022-5.00008-7

Wu, W., Emerton, R., Duan, Q., Wood, A. W., Wetterhall, F., & Robertson, D. E. (2020). Ensemble flood forecasting: Current status and future opportunities. WIREs Water, 7(3). https://doi.org/10.1002/wat2.1432

Xie, P., Arkin, P. A., & Janowiak, J. E. (2007). CMAP: The CPC merged analysis of precipitation. In Measuring precipitation from space. In V. Levizzani, P. Bauer, & F. J. Turk (Eds.) (Vol. 28, pp. 319–328). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5835-6_25

Xu, B., Zhong, P., Lu, Q., Zhu, F., Huang, X., Ma, Y., & Fu, J. (2020). Multiojective stochastic programming with recourses for real-time flood water conservation of a multireservoir system under uncertain forecasts. Journal of Hydrology, 590, 125513. https://doi.org/10.1016/j.jhydrol.2020.125513

Xu, J., Anctil, F., & Boucher, M.-A. (2022). Exploring hydrologic post-processing of ensemble streamflow forecasts based on affine kernel dressing and non-dominated sorting genetic algorithm II. Hydrology and Earth System Sciences, 26(4), 1001–1017. https://doi.org/10.5194/hess-26-1001-2022

Xu, W., Fletcher, T. D., Burns, M. J., & Chenqui, F. (2020). Real time control of rainwater harvesting systems: The benefits of increasing rainfall forecast window. Water Resources Research, 56(9). https://doi.org/10.1029/2020WR027856

Yuan, X., Ji, P., Wang, L., Liang, X., Yang, K., Ye, A., et al. (2018). High-resolution land surface modeling of hydrological changes over the sanjiangyuan region in the eastern Tibetan plateau: 1. Model development and evaluation. Journal of Advances in Modeling Earth Systems, 10(11), 2806–2828. https://doi.org/10.1029/2018MS001412

Zarekarizi, M., Yan, H., Ahmadalipour, A., & Moradkhani, H. (2021). A probabilistic framework for agricultural drought forecasting using the ensemble data assimilation and Bayesian multivariate modeling. In H. Wu, D. P. Lettenmaier, Q. Tang, & P. J. Ward (Eds.), Geophysical monograph series (1st ed., pp. 147–164). Wiley. https://doi.org/10.1002/978111927339.ch8

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E., & Ramos, M.-H. (2017). How suitable is quantile mapping for postprocessing GCM precipitation forecasts? Journal of Climate, 30(9), 3185–3196. https://doi.org/10.1175/JCLI-D-16-0652.1

Zhao, T., Wang, Q. J., & Schepen, A. (2019). A Bayesian modelling approach to forecasting short-term reference crop evapotranspiration from GCM outputs. Agricultural and Forest Meteorology, 269(270), 88–101. https://doi.org/10.1016/j.agrformet.2019.02.003

Zhao, T., Yang, D., Cai, X., Zhao, J., & Wang, H. (2012). Identifying effective forecast horizon for real-time reservoir operation under a limited inflow forecast. Water Resources Research, 48(1). https://doi.org/10.1029/2011WR010623