Environmental Modelling & Software 92 (2017) 152–162

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Improvement of rank histograms for verifying the reliability of extreme event ensemble forecasts

Jing Xu [a, c], Aizhong Ye [a, b, *], Qingyun Duan [a, b], Feng Ma [a, b], Zheng Zhou [a, b]

[a] State Key Laboratory of Earth Surface and Ecological Resources, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[b] Joint Center for Global Change Studies, Beijing 100875, China
[c] Dept. of Civil and Water Engineering, Université Laval, 1065 avenue de la Médecine, Quebec, Canada

ARTICLE INFO

Article history:
Received 22 April 2016
Received in revised form
16 February 2017
Accepted 22 February 2017

Keywords:
Rpolar diagram
Extreme events
Ensemble forecast verification
Partial reliability

ABSTRACT

Ensemble forecasting is becoming increasingly popular as a hydrological forecasting tool because of its advantages of not only predicting the most likely outcome of a hydrological event, but also providing related uncertainty information. Rank histogram is one of the most widely used metrics for verifying the reliability of ensemble forecasts. This study proposed an improved rank histogram method, the Rank Polar Diagram (Rpolar diagram), for evaluating the reliability of ensemble forecasts of extreme events. The conventional rank histogram provided a simple evaluation of the reliability of an ensemble forecast, which could not differentiate the reliability of the forecasts of extreme events such as heavy storms, high flows and low flows. Rpolar diagrams were able to verify not only the overall reliability but also the partial reliability over different flow intervals, including extremes. In Rpolar diagram, forecast intervals could be set according to user preference or uniform intervals automatically. This study evaluated the effectiveness of the Rpolar diagram using two typical sets of simulation ensembles and actual streamflow/precipitation ensembles. Both streamflow and precipitation application results exhibited the suitability of the Rpolar diagrams for verifying the reliability of extreme events.

1. Introduction

Accurate and reliable forecasting of hydrological events, especially extreme events such as heavy storms, droughts, and floods, was imperative for the protection of socio-economy and human lives. Ensemble forecasting (Van Steenbergen et al., 2012; Gal et al., 2014) has become an increasingly popular tool for the forecasting of hydrological extremes. However, how can one verify the quality of an ensemble forecast properly? Generally speaking, a good ensemble forecast was supposed to have following: 1) equal likelihood: each ensemble member having an equal likelihood of occurrence, 2) superiority: the ensemble mean being superior to any single forecast when evaluated over a long verification period, 3) high spread-skill relation: the ensemble spread being proper and distribution being sharp, and 4) high reliability: forecast probability matching observed frequency.

Various verification metrics have been proposed over the years to evaluate the performance of an ensemble forecast (Murphy and Winkler, 1987; Nash and Sutcliffe, 1970; Jolliffe and Stephenson, 2008) and single-valued forecasts (Dawson et al., 2007), with aforementioned properties measured separately by different verification metrics (Potts et al., 1996; Pushpalatha et al., 2012). Table 1 listed how those properties were measured by various verification metrics.

The large numbers of verification metrics listed above indicated that the goodness of ensemble forecasts could not be evaluated straightforwardly with a single measure. In other words, users would need to verify the ensemble forecasts with different metrics to obtain a complete picture of its performance. For example, we could use a number of verification metrics that measure reliability, such as the rank histograms (Talagrand et al., 1997), predictive quantile-to-quantile plots (Laio and Tamea, 2007; Biondi and De Luca, 2013), reliability diagrams (Hamill, 1997), among others. For other properties such as resolution, sharpness, discrimination, one needed to use other metrics, e.g., the relative operating characteristic (*ROC*) (Swets, 1973), the decomposition of the brier score (*BS*) into the resolution term (Murphy, 1973), or mean continuous rank probability scores. This paper focused on one of these metrics, the rank histogram, and discussed how rank histograms could be

* Corresponding author. State Key Laboratory of Earth Surface and Ecological Resources, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China.
E-mail address: azye@bnu.edu.cn (A. Ye).

http://dx.doi.org/10.1016/j.envsoft.2017.02.024
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

Table 1
Studies with the certain criteria to evaluate different attributes of forecast quality (Brown et al., 2010).

Aspects	Property	Definition	Verification metrics & Efficiency criteria	References
Ensemble mean	Bias	Measure the difference between the forecast and the observed values on average.	Mean Error (ME) Mean Absolute Error (MAE) Root Mean Square Error (RMSE) Mean Continuous Rank Probability Score	Willmott and Matsuura, 2005; Singh et al., 2014.
	Accuracy	The correlation between the forecast and the observations on average.	Brier Score (BS)	Brier, 1950; Stephenson et al., 2008.
	Correlation	Degree of the linear relationship between the forecast and the observations on average.	Pearson Correlation Coefficient (PCC) Spearman Rank Correlation (SRC) Kendall Correlation Coefficient	Benesty et al., 2009; Zar, 1998; Abdi, 2007.
	Skill	Extent degree of whether the forecast system is better or worse than the benchmark (e.g., climatology).	Mean Continuous Rank Probability Skill Score (MCRPSS) Brier Skill Score Equitable Threat Score	Hersbach, 2000; Weigel et al., 2007
Probability	Reliability	Degree of correlation between the forecast probability and the observed frequency.	Reliability Diagram Rank Histogram Predictive Quantile-Quantile Plot	Hamill, 1997; Hamill, 2001; Maraun, 2013.
	Resolution	An attribute that measures how well the observations are sorted under different frequency distributions.	Mean CRPS Resolution Brier Score Resolution Relative Operating Characteristic	Beck et al., 1986; Jolliffe and Stephenson, 2008.
	Discrimination	Degree of forecast ability to discriminate between events and non-events.	Relative Operating Characteristic Score	Mason and Graham, 1999.
Spread-skill relation	Sharpness	An attribute that measures the tendency to predict with zero and one.	Forecast Frequency Histogram	Hoffmann et al., 1990; Sircombe, 2004.

improved so it could be used to verify the reliability of ensemble forecasting of extreme events.

Wilks (2011) explained how an ensemble forecast was used to generate a probabilistic forecast through the following example: if 20 of 50 ensemble members forecasted rain, then the probability of rain would be 40%. The premise of this example was that the ensemble was reliable, which implied that both the ensemble forecast and the observation were sampled randomly from the same probability distribution.

It was common to diagnose the reliability of ensemble forecasts from the shape of rank histograms, predictive *QQ* plots or reliability diagrams. Rank histogram is also known as the “Talagrand diagram” (Talagrand et al., 1997). Wilks (2011) suggested using the rank histogram to evaluate the ensemble spread of the forecast value. It checks whether the observed probability distribution is well represented by the ensembles. Arrange the N ensemble members in an increasing order from lowest to highest and we will obtain $(N + 1)$ bins that the verifying observation could fit into, including the two extremes. Ideally speaking, each ensemble member represents an equally likely scenario to contain the observed value in the corresponding bin. A flat rank histogram manifests ensemble spread about right to represent forecast uncertainty. Some other situations may occur as well, such as a U-shaped rank histogram represents the distribution of ensemble members is too small that many observations falling outside the extremes; A dome-shaped rank histogram shows the distribution of ensemble members is too large that most observations falling near the center of the ensemble; An asymmetric rank histogram denotes the bias in the ensemble. However, Hamill (2001) indicated that uncritical use of these tools might lead to an inaccurate evaluation of the reliability of ensemble forecasts. An important theory claimed that a uniform rank histogram was a necessary but insufficient criterion for determining whether an ensemble was reliable. A uniform rank histogram provided no guarantee that the ensemble was reliable at each point used to populate the histogram. On the other hand, the forecast reliability of extreme events was crucial, such as the evaluation of the partial reliability of flood, low-flow, heavy precipitation and others.

This study proposed an improved verification method called the Rank Polar Diagram (Rpolar diagram) for evaluating the reliability of ensemble forecasts of extreme events. Section 2 introduced the rank polar diagram and illustrated its effectiveness through several typical examples. Section 3 presented the results of Rpolar diagrams from actual forecasting applications in different hydroclimate regimes. Section 4 provided concluding remarks.

2. Rank polar diagram (Rpolar Diagram)

2.1. Introduction of rank polar diagram

Wilks (2011) suggested that rank histogram could be used to evaluate the reliability of ensemble forecasts. The rank histogram verified whether the observed probability distribution was represented well by the ensembles and diagnoses errors within its mean and spread (Talagrand et al., 1997; Hersbach, 2000; Yuan et al., 2013). However, an approximately uniform rank histogram was a necessary, rather than sufficient, condition for a reliable forecast or a perfect-model context. In other words, the ensemble forecast that was reliable at each point would generate a uniform rank histogram; however, a flat rank histogram nevertheless could be generated from unreliable ensembles (Hamill, 2001).

The concern when verifying the reliability of ensemble forecasts is higher for extreme events, such as heavy storms, floods, droughts and low-flow discharge (Dong et al., 2013), extreme heat, etc. Therefore, it is important to ensure ensemble reliability at each point. An improved verification method utilizing a Rank Polar (Rpolar) Diagram was suggested in this paper. The Rpolar Diagram could be a useful tool for evaluating the partial reliability of extreme events in ensemble forecasts. The Rpolar Diagram was constructed as follows: 1) given the number of observations n, and the number of ensemble members m, sort the observed data in descending order, and divide the data into 10 quantiles; 2) for each quantile, obtain the rank histogram of the corresponding ensemble forecasts; 3) compute the root-mean-squares error ($RMSE_i$) of the rank values versus the perfect rank (in a perfect-model context) value of $1/(m+1)$ for quantile i, $i = 1,2, \ldots, 10$, and then compute *mRMSE*, which is the weighted average of $RMSE_i$ (i = 1, 2 ..., 10), noted that equal weights of 1/10 are used here; 4) draw a red circle representing the perfect rank (with a radius value of 1/(m+1)), and then draw the dashed radial lines separating the 10 quantiles; 5) for each quantile, plot the portion of the rank histogram whose values are larger than 1/(m+1) in red and the portion whose values are less than 1/(m+1) in cyan; 6) draw the dased circles with percentages (2%, 4%, 6%, 8%) that represent the observed frequencies; 7) draw the $RMSE_i$ divided by 100 values for each quantile, (marked by "x") in order to match the polar axis. Noted that the root mean square error ($RMSE_i$, i = 1, 2, ..., 10) is computed as follows:

$$RMSE_i = 100 \bullet \sqrt{\frac{1}{m+1}\sum_{k=1}^{m+1}\left(R_k - \frac{1}{m+1}\right)^2} \tag{1}$$

where i represents the sequence number of the observational frequency interval, m is the number of ensemble members, and k represents the sequence number of each ensemble member. R_k represents the rank values of each ensemble member in the corresponding intervals. $RMSE_i$ varies between 0 (perfect reliability) and $+\infty$ (worst reliability).

We created a synthetic case to illustrate the Rpolar Diagram. In this case, we assumed that there are 100,000 observations, which were divided into 10 quantiles, with each quantile containing 10,000 observations. Further, we assumed that all observations are drawn from Gaussian distributions, $N(\mu_i,\sigma_i)$, i = 1, ..., 10, where μ_i and σ_i are the mean and standard deviation of the ith quantile. Synthetic ensemble forecasts for the ith quantile were generated as follows. Assumed that the ensemble size was 25. For each observation in the ith quantile, 25 random values were sampled from $N(\mu_i+\Delta\mu_i,\sigma_i+\Delta\sigma_i)$, where $\Delta\mu_i$ and $\Delta\sigma_i$ were perturbations to μ_i and σ_i. Thus, for 10,000 observations in the ith quantile, 10,000 × 25 random values were sampled from $N(\mu_i+\Delta\mu_i,\sigma_i+\Delta\sigma_i)$. By changing the values of $\Delta\mu_i$ and $\Delta\sigma_i$, we created various kinds of unreliable ensemble forecasts for different quantiles (Table 2). Fig. 1a also provided a schematic diagram of ensemble forecasts with a particular set of {$\Delta\mu_i$ and $\Delta\sigma_i$, i = 1, 2, ..., 10}, with the resulting rank histogram shown in Fig. 1b. The overall rank histogram seemed to be quite uniform except in the 1st to 26th quantiles.

Fig. 1c shows the Rpolar Diagram for the above ensemble forecasts. If both observations and ensemble forecasts are sampled from the same distribution (both $\Delta\mu_i$ and $\Delta\sigma_i$ are equal to 0), the ensemble forecasts should be perfectly reliable according to the definition of perfect reliability. This shows up in the 90% and 100% quantiles in Fig. 1c, where the rank histograms lie close to the red line (i.e., perfect rank). If either or both $\Delta\mu_i$ and $\Delta\sigma_i$ are not equal to 0, then the ensemble forecasts are going to be unreliable, as shown in the other 8 quantiles. A U-shaped rank histogram is showed in the 1st and 2nd quantiles in Fig. 1c, the ensemble samples are from a distribution with a lack of variability when $\Delta\sigma_i$ is less than 0 and $\Delta\mu_i$ is equal to zero. An excess of variability in the ensemble would overpopulate the middle ranks and result in a dome-shaped Rpolar diagram in the 3rd and 4th quantiles in Fig. 1c ($\Delta\sigma_i$ >0 and $\Delta\mu_i = 0$).

Table 2
Corresponding perturbations to sample mean and standard deviation and type of error for each frequency intervals.

Sample	No. of quantiles	$\Delta\mu_i$	$\Delta\sigma_i$	Type of error
1	10-20%	=0	<0	Lack of variability
2	30-40%	=0	>0	Excess of variability
3	50-60%	>0	=0	Under-forecasting
4	70-80%	<0	=0	Over-forecasting
5	90-100%	=0	=0	No error

If $\Delta\mu_i$ is greater than 0, the rank histograms demonstrate under-forecasting in the 5th and 6th quantiles in Fig. 1c. If $\Delta\mu_i$ is less than 0, the rank histograms demonstrate over-forecasting in the 7th and 8th quantiles in Fig. 1c.

Note that a perfectly reliable ensemble forecast is a special case in which the rank histograms for all quantiles coincide with the red circle (i.e., perfect rank) and the $RMSE_i$ values are all zeros at the circle center.

The choice of quantile number should be dependent on sample size and user demand. If the sample size is large enough (e.g., >10000), the quantile number might be greater than 10. If the sample size is very small (e.g., <100), the quantile number would be less than 10. The Rpolar Diagram would be a rank histogram if the quantile number equals 1. The choice of 10 quantiles is optimum scheme in order to be used easily. Besides, the sample size in each quantile may be different (e.g., 1st: 0–5%, 2nd:5–20%,... ...) .

2.2. Comparison between rank polar diagram and conventional rank histogram

Rpolar Diagram is also well suited to illustrate the unsettled issues given by Hamill (2001), which showed that an ensemble forecast could only have whole reliability, but partial unreliability.

Truth was drawn randomly from a standard normal distribution N (0, 1), and a 25-member ensemble was sampled with equal likelihood from one of three probability distributions: N (−0.5, 1), N (0.5, 1) and N (0, 1.3) (Fig. 2a). A relatively flat rank histogram was generated (Fig. 2b) despite none of the ensembles sampled from the same probability distribution representing truth. The ensemble was clearly unreliable in this example because of the positive/negative conditional biases and an excessive variability, so the resulting rank histogram is a misinterpretation.

Fig. 2a shows the observation and 3 ensemble forecasts which are sampled from three slightly different distributions. If one verifies the 3 ensemble forecasts together, the ensemble looks perfectly reliable, as shown by the rank histogram (Fig. 2b). Hamill (2001) argued that a uniform rank histogram does not necessarily indicate the reliability of the ensemble by drawing conclusions from a flat rank histogram populated with ensembles from three different probability distributions than that from which the observation is drawn. Here, the Rpolar diagram (Fig. 2c) can evaluate the forecast quality (or partial reliability) at each sample point more clearly than the rank histogram.

Fig. 2c indicates that the Rpolar diagram can reveal the unreliability of the ensemble. The partial reliability at each frequency range in the ensemble can be evaluated. For example, the probability distribution of N (0.5, 1) with a relatively large mean of 0.5 (Fig. 2a, green line) results in asymmetrically shaped ranks with smaller ranks close to member 25 at 8th to 10th quantiles. This result indicates over-forecasting in a low-frequency interval and corresponds to an N (−0.5, 1) distribution (Fig. 2a, blue line). The relative ensemble mean of −0.5 results in excessively populated ranks close to member 25 and indicates under-reliability in high-frequency intervals (1st to 3rd quantiles). Dome-shaped Rpolar diagrams (Fig. 2c) with excess variability are generated in middle frequency intervals (4th to 7th quantiles) when the ensemble is sampled from a distribution of N (0, 1.3) with a relatively large standard deviation of 1.3 (Fig. 2a, orange line).

The resulting $RMSE_i$ values are 8.04, 4.50, 3.64, 3.10, 2.79, 2.84, 3.11, 3.56, 4.58 and 8.02 (Fig. 2c) corresponding to 1st to 10th quantiles, respectively. The *mRMSE* value of 4.42 is much higher than the original *RMSE* value of 0.77 (Fig. 2b), indicating unreliability in the ensemble. Therefore, the Rpolar diagram can interpret the qualities of the ensemble properly.

Forecast performance typically was evaluated separately for

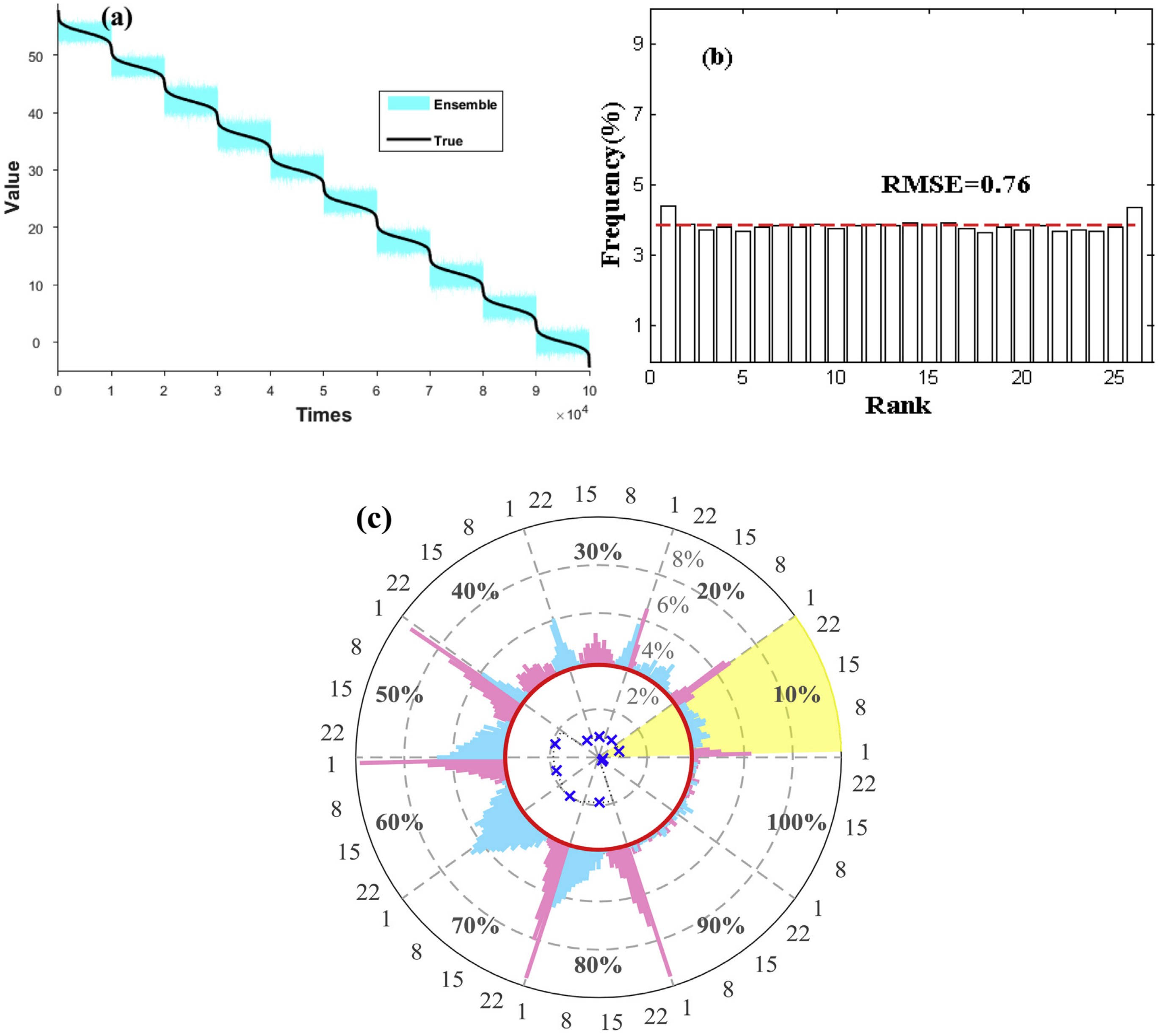

Fig. 1. (a) Schematic diagrams of five different ensemble samples. The observation is sampled randomly with a standard normal distribution $N\,(6(10\text{-}i), 1)$ ($i = 1$ to 10), and five 25-member ensemble samples are drawn from $N\,(6(10\text{-}i), 0.7)$ ($i = 1$ and 2), $N\,(6(10\text{-}i), 1.5)$ ($i = 3$ and 4), $N\,(6(10\text{-}i)+0.5, 1)$ ($i = 5$ and 6), $N\,(6(10\text{-}i)\text{-}0.5, 1)$ ($i = 7$ and 8) and $N\,(6(10\text{-}i), 1)$ ($i = 9$ and 10), where i represents the sequence number of the frequency intervals. Note that the same amounts are increased in each 10000 times. (b) The corresponding rank histogram, where x-axis denotes the ensemble members, y-axis denotes the frequency of observation in corresponding ensemble forecast. (c) The Rpolar diagram. The dashed radial lines separate the 10 quantiles (10%–100%, in descending order) that the observed data were divided into. The red circle shows the perfect rank. The dashed circles different percentages (2%, 4%, 6%, 8%) represent the frequencies. The root-mean-squares error ($RMSE_i$) vales for each quantile are marked by blue “x”. Furthermore, for each quantile, the red portions represent the values are larger than 1/(m+1), while the cyan portions are for the values are less than 1/(m+1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

each regime (i.e., different areas, seasons, models, etc). However, Hamill (2001) suggested an indistinguishable problem within the rank histogram and argued that a rank histogram of any shape may be generated in a variety of ways. Hamill (2001) discussed that two types of unreliability (conditional bias and under variability) might lead to similar U-shaped rank histograms. The Rpolar diagram can solve this problem and distinguish among different types of unreliability or errors in the ensemble.

The truth is sampled randomly from a standard normal distribution $N\,(0, 1)$. Two pairs of 25-member ensembles are selected as follows (Fig. 3a and b): one is sampled with equal likelihood from one of two different probability distributions other than the truth. The two distributions include N (−1, 1) and N (1, 1); another ensemble is sampled from an under-variable population via a distribution of $N\,(0, 0.42)$. The first pair of ensembles in this example contains a combination of conditional biases, while the second pair is sampled with a lack of variability. However, the rank histograms shown in Fig. 3c and d are too similar for diagnosing different types of mean and spread errors in the ensemble.

The Rpolar diagrams in Fig. 3e and f are completely different, implying two types of unreliability in the corresponding ensemble. The asymmetrically shaped ranks with smaller extreme ranks indicate a positive bias in the 100% high-frequency interval (10th quantile). The overpopulation of large ranks implies a negative bias

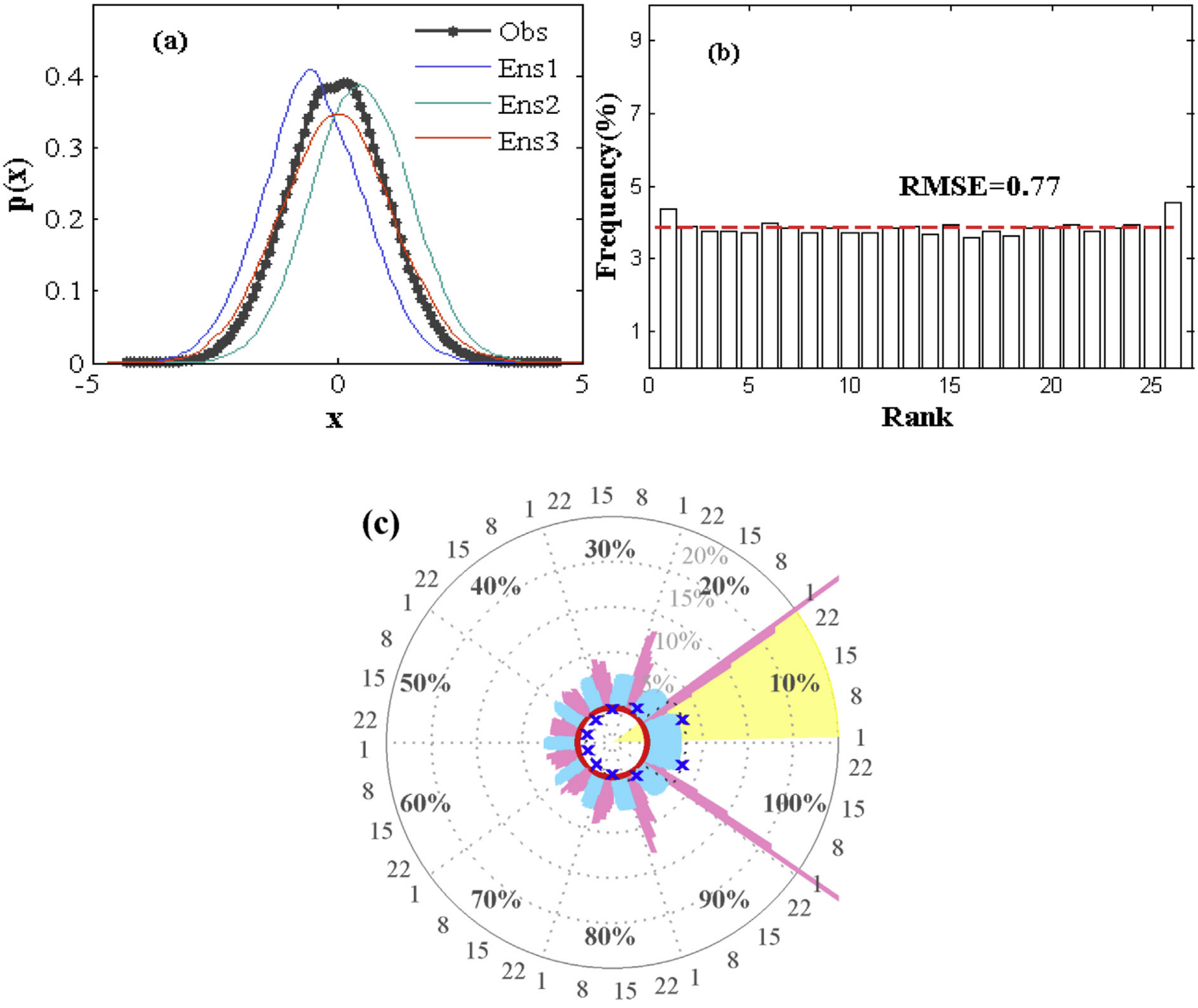

Fig. 2. (a) Probability distributions from which verification and ensemble are sampled. The observation is sampled randomly with a standard normal distribution $N\,(0,1)$, and three 25-member ensemble samples are drawn from $N\,(-0.5,1)$, $N\,(0.5,1)$ and $N\,(0,1.3)$. (b) Resulting rank histogram corresponding to (a). (c) Rpolar diagram corresponding to (a). The rank of the verification and ensemble are tallied 30000 times total.

in the ensemble in the 10% low-frequency interval (1st quantile). The unreliability is also displayed at 2nd to 9th quantile. However, a positive bias excessively populates the small extreme ranks in high-frequency intervals from 80% to 100% (Fig. 3f). A negative bias overpopulates the large extreme ranks in low-frequency intervals from 10% to 30%. Dome-shaped ranks (Fig. 3f) indicate over-variability in the second ensemble for the middle frequency intervals from 40% to 70%.

The *RMSE* values (2.17 and 2.41) in the rank histograms in Fig. 3c and d, are indistinguishable. Table 3 shows that the values of *mRMSE* are 5.55 and 8.42, respectively. The different $RMSE_i$ in Rpolar diagrams can distinguish different types of unreliability in the different quantiles.

3. Applications in ensemble forecast verification

Ensemble forecasts generally are considered to perform differently under different weather regimes. Simulated and observed daily streamflow/precipitation data indicate that these ensemble forecasts encompass divergent regimes. Extreme events (such as flood, drought or low-flow, heavy precipitation and extreme heat, etc.) within a real hydrological ensemble forecast are important integral components of any weather regime.

3.1. *Streamflow cases*

The simulated daily streamflow data and corresponding observation used in this study were obtained from the database of the Second Workshop on Model Parameter Estimation Experiment (MOPEX) (Duan et al., 2006), which involved 12 basins from the Southeastern United States and covered the period from 1962 to 1997. The simulated streamflow predictions by seven hydrological models from 3 of the 12 basins have been post-processed to generate ensemble streamflow predictions (Ye et al., 2014). Those models are the Gr4j (gr4j), Isba (isba), Noah (noa), Sacramento (sac), SWAP (swap), Simple Water Balance (swb) and VIC models (vic). Some basic information of the three river basins is shown in Table 4. Here, we choose three typical streamflow examples to demonstrate how Rpolar diagrams can evaluate partial reliability of different frequency intervals in the ensemble, especially extreme events (i.e., floods).

3.1.1. *ExampleⅠ: partial reliability of different frequency intervals accurately of ensemble streamflow simulations*

Both the resulting rank histogram (Fig. 4a) and Rpolar diagram (Fig. 4b) are displayed for comparison. Note that the entire period is from 1987 to 1997 for the two figures. Fig. 4c shows a hydrograph of

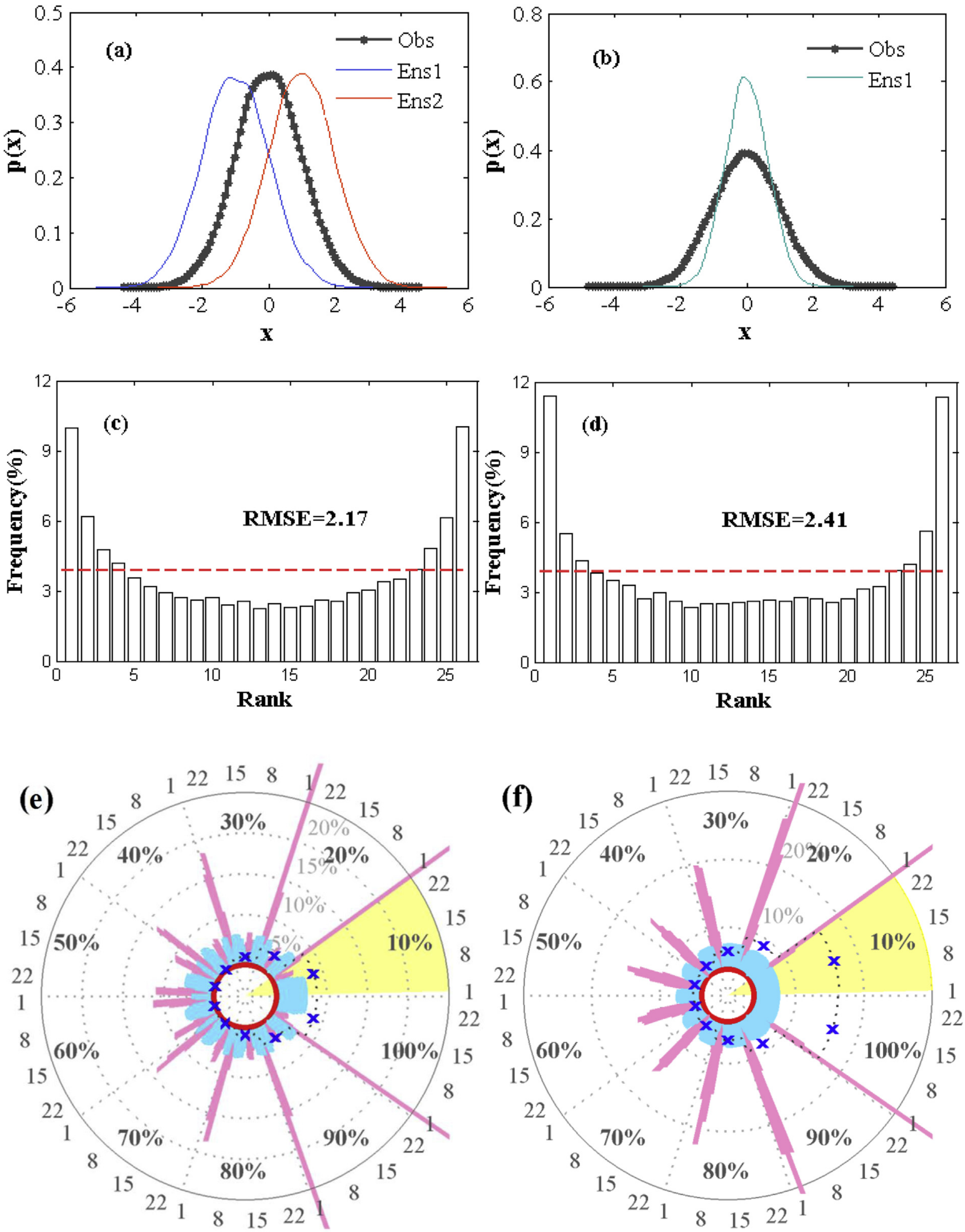

Fig. 3. (a)(b) Probability distributions from which verification and ensemble are sampled. The observation is sampled randomly with a standard normal distribution $N(0, 1)$, and the first 25-member ensemble samples are drawn from $N(-1, 1)$ and $N(1, 1)$. The second ensemble is sampled with $N(0, 0.49)$. (c)(d) Rank histograms corresponding to (a) and (b), respectively. (e)(f) Rpolar diagram corresponding to (a) and (b), respectively. The rank of the verification and ensemble are tallied 20000 times in each ensemble sample.

post-processed ensemble streamflow simulations for B1 basins (Table 4) with relatively low latitude for the Gr4j model. The chosen period in Fig. 4c is from January 1, 1992, to December 31, 1992, to illustrate a schematic hydrograph. The post-processed ensemble streamflow simulation data we used have lead time of 14 days, while, in this study, we only chose the 1st lead time.

Table 3
The values of *RMSE* for two different ensembles.

Criteria		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	*mRMSE*
RMSEi	1#	8.78	6.15	4.81	4.21	3.83	3.78	4.20	4.77	6.09	8.92	**5.55**
	2#	16.06	8.81	6.43	5.56	5.00	5.08	5.52	6.39	8.97	16.39	**8.42**

Table 4
River basin information.

Basin ID	USGS ID	Lon.	Lat.	Area(km^2)	Station name
B1	08172000	−97.6506	29.6661	2170	San Marcos River at Luling, TX
B2	01608500	−94.5661	37.2456	3810	S Branch Potomac River near Springfield, WV
B3	01668000	−77.5181	37.2456	4134	Rappahannock River near Fredericksbrug, VA

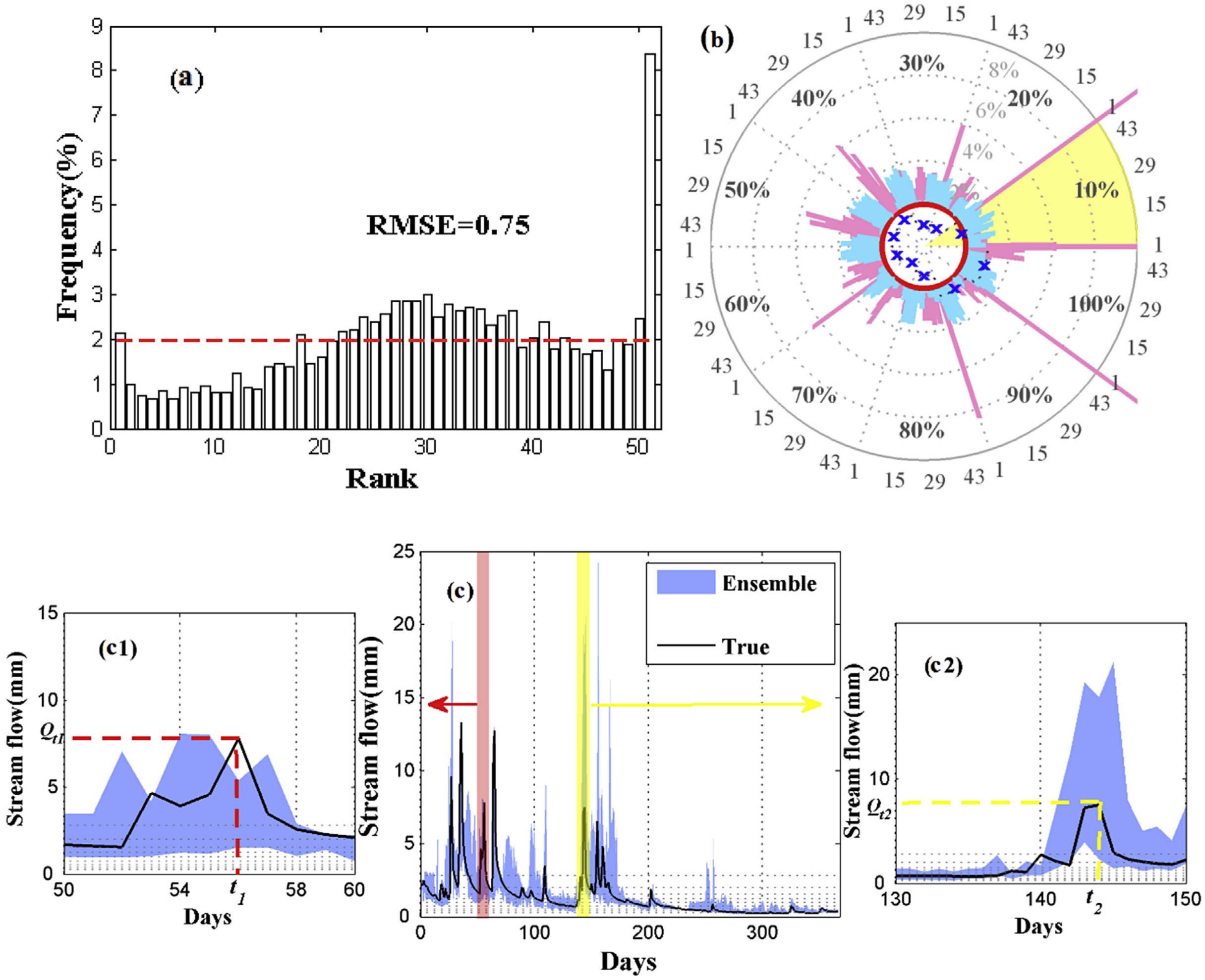

Fig. 4. (a) Rank histogram from 1987 to 1997. The horizontal red dashed lines indicate perfect uniformity. (b) Corresponding Rpolar diagram. The perfect uniformity is shown as a red circle, while the blue crosses in each sector domain represent the $RMSE_i$. (c) Hydrograph for post-processed ensemble streamflow simulations on lead day 1 from January 1, 1992, to December 31, 1992, for the B1 river basin using the Gr4j model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Two major features can be summarized according to the resulting rank histogram (Fig. 4a): (1) the dome-shaped rank histogram with peaks located in the center indicates an excess of variability in the ensemble and (2) the excessive population of the right extreme ranks reveal a negative bias in the ensemble. The results for whether the biases occurred within the high discharge cannot be acquired through a rank histogram alone.

A corresponding Rpolar diagram is plotted in Fig. 4b. Note that the yellow fan areas represent the high discharge interval. We found that (1) the smallest and largest ranks are high indicate both positive (over-forecasting) and negative biases exist in the flood high ensemble simulations in the low-frequency interval (10–20%); the polar ranks appear U-shaped at 1st and 2nd quantiles. Therefore, the verification result reflects under-forecasting and excessive variability marked with the red/yellow band in Fig. 4c. (2) the overpopulation of the middle ranks (over-variability) appear in middle-frequency intervals from 30 to 60%; and (3) a negative bias (under-forecasting) excessively populates the large extreme ranks in high-frequency intervals from 80 to 100%. A significant number of ensemble members exceed the ensemble upper uncertainty range in the high discharge interval, which indicates huge risk. Therefore, this streamflow example can be considered an unreliable ensemble based on the above analysis. The Rpolar diagram can interpret the forecast performance visualized by the hydrograph better than the rank histogram. The values of $RMSE_i$ are 1.87, 1.02, 1.01, 1.53, 1.49, 1.30, 0.93, 1.39, 2.46 and 2.99 (Fig. 4b) for the period from 1987 to 1997 for the B1 river basin according to the Gr4j model. The *mRMSE* value of 1.60 is much larger than the original *RMSE* value of 0.75, indicating unreliability in the ensemble. Therefore, the Rpolar diagram can interpret partial reliability of different frequency intervals accurately.

The overall impression given by the hydrograph (Fig. 4c) is that there is an under-confidence in the ensemble simulations. The ensemble generates a "band" (Fig. 4c, pink area), the size of which is too small to contain the truth (black line). Conditional biases exist within the high flood intervals. For example, under-forecasting is visualized, indicating that the observed streamflow rises higher than the ensemble flood peak prediction. Similarly, over-forecasting and excessive variability (yellow area) are demonstrated. Note that 10 different frequency quantiles based on observations are plotted as black dashed lines (Fig. 4b).

3.1.2. ExampleII: distinguish the forecast performance of different ensemble streamflow simulations

Fig. 5 provides another example in which the Rpolar diagram is used to distinguish the forecast performance of different ensemble streamflow simulations. We chose the Sacramento model data in the B2 and B3 river basins as the streamflow data used in this example. The hydro-climate regimes are clearly different in the view of the complementary information of the B2 and B3 basins (Table 4). Fig. 5a and b indicate that the shapes of the two rank histograms are the same with *RMSE* values of 0.56. The rank histograms are indistinguishable and useless for verifying different reliabilities in each weather regime in this case. The $RMSE_i$ (*mRMSE*) values with a weight of one-tenth are also demonstrated in Table 5.

The Rpolar diagrams (Fig. 5c and d) demonstrate explicit results. We found that the dome-shaped polar ranks appear in almost all of the domain sectors of the two Rpolar diagrams. However, the 10% red rank indicates an under-forecasting in the ensemble for the high discharge (yellow fan area) in the prior Rpolar diagram (Fig. 5c). In contrast, the red ranks of closing member 1 in the yellow fan area of Fig. 5d reveal over-forecasting in the 10% interval. Although the sharpness is relatively low, few ensemble members exceed the ensemble upper uncertainty range in the high discharge interval.

3.1.3. Example III:perfect ensemble streamflow forecast with high reliability

The merit of Rpolar diagram is which can show more uncertainly information of ensemble forecast than rank histogram. The Rpolar diagram can check not only whole reliability but also partial reliability. Fig. 6 shows the rank histogram and Rpolar diagram in allusion to a perfect reliable ensemble streamflow simulation. The perfect reliable ensemble streamflow simulation was produced randomly with uniform distribution on each observed daily streamflow among 1961 to 1998 in the B3 basin. There are four steps to generate the perfect reliable ensemble daily streamflow simulation in each day: 1) Produce 25 uniform distributed random numbers $a(1 \ldots 25)$, 2) Sort the 25 random numbers in descending order, 3) Produce one random integer number k among 1 to 26, 4) the observed streamflow minus $a(1, \ldots, k\text{-}1)$ and add $a(k, \ldots, 25)$ if k is not equal to 1 and 26, the observed streamflow add $a(1, \ldots, 25)$ if k is equal to 1, the observed streamflow minus $a(1, \ldots, 25)$ if k is equal to 26.

Fig. 6a is a uniform rank reveals the ensemble simulation is reliable. The resulting $RMSE_i$ values are 0.5789, 0.3812, 0.5725, 0.5974, 0.4705, 0.5180, 0.5417, 0.5021, 0.6105, and 0.5053 (Fig. 6b) corresponding to at 1st to 10th quantiles, respectively. The uniform rank values and small $RMSE_i$ show the ensemble forecast is partial reliable (Fig. 6a). The Rpolar diagram can show a reliable forecast to be reliable on ensemble streamflow simulation.

3.2. Precipitation case

In terms of precipitation, the daily precipitation ensemble forecast data used in this study were generated through post-processing of the single-value Global Forecast System (GFS) precipitation forecasts over the Huai River basin in China over 28 years from 1981 to 2008. The post-processing method used in this study was the revised Ensemble Predicted Preprocessor (EPP), originally developed at the Hydrology Laboratory of the National Weather Service (Schaake et al., 2007; Liu et al., 2013). The daily precipitation ensemble forecasts contained 28 members with a lead-time of 14 days. The corresponding observed daily precipitation observations were collected from 187 weather stations located in and near the river basin. The annual precipitation amount ranging from 634 to 1130 mm/yr, decreased from the Southwest to the Northeast.

Fig. 7a shows the precipitation chart of the ensemble precipitation simulation in day 1 for the period (1981–2008). The resulting rank histogram (Fig. 7b) and Rpolar diagram (Fig. 7c) are demonstrated for comparison. It is worth noting that there may be no precipitation in the arid season from October to March when verifying the reliability of the ensemble precipitation forecasts. The verification observation including the two extremes will fit into obtained $(N + 1)$ bins when arranging N ensemble members into an increasing order. For example, if the observation and M $(M < N)$ ensemble members are equal to zero, the ranks will be over-populated in the $(M + 1)^{th}$ bin. However, the prior $(M + 1)$ bins have equally likely scenario to contain the observed value. Therefore, special rules are needed for assigning ranks when a large number of ensemble members have the same value (i.e., zero). The cumulative amounts in the $(M + 1)^{th}$ bin are supposed to be divided into the former M bins uniformly.

Fig. 7a indicates that the size of the "band" (light blue area) that is comprised of the ensemble is relatively suitable for the truth (grey line) and will give users more confidence in decision making the ensembles with a larger spread. However, conditional biases also exist in low-frequency intervals (i.e., heavy precipitation). For instance, the observation exceeds the ensemble upper uncertainty range (Fig. 7a, above the red line), reveals under-forecasting.

The resulting rank histogram is shown in Fig. 7b. We found that (1) the ranks are nearly uniform and (2) the right-most ranks are relatively high, suggesting that there is a negative bias (under-forecasting) in the ensemble. The rank histogram generally suggests that the ensemble precipitation simulation is reliable.

Nevertheless, opposite conclusions can be obtained by the Rpolar diagram (Fig. 7c). Two major results are found: (1) the precipitation probability is approximately 30%, indicating an approximately 70% probability of no rain. Most ensemble members are equal to or greater than zero in the low-frequency intervals (from 40% to 100%), as with the corresponding observation. The ranks of prior $(M + 1)$ bins in each interval are distributed uniformly determined using the rank assigning method discussed above. The large extreme ranks, such as the (*N-M*) bins, are under-

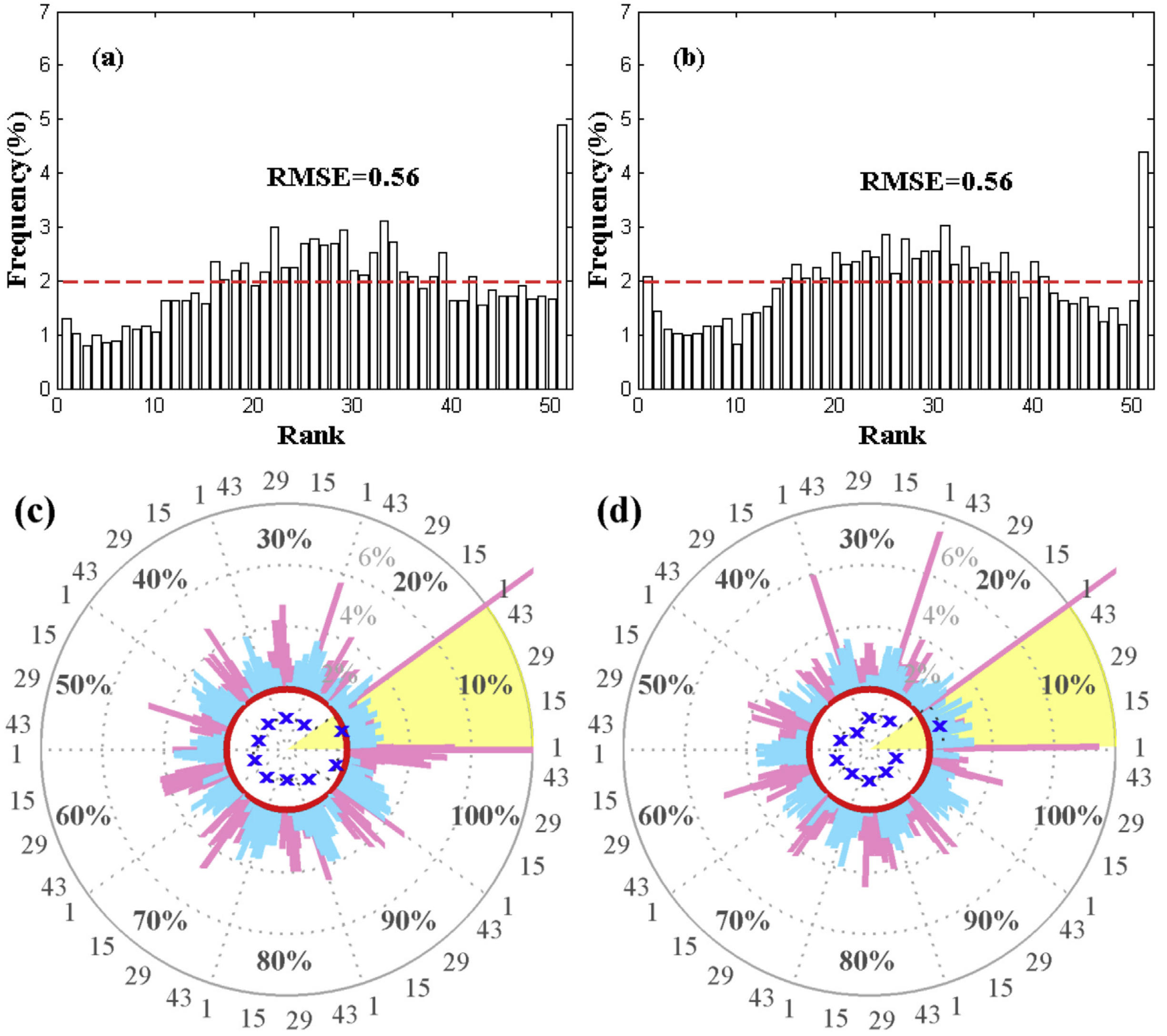

Fig. 5. (a)(b) Rank histograms for post-processed ensemble streamflow simulations on day 1 for the B2 and B3 river basins using the Sacramento model. (c)(d) Corresponding Rpolar diagrams.

Table 5
The values of *RMSE* for the B2 and B3 river basins from 1987 to 1997 using the Sacramento model.

Criteria		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	*mRMSE*
RMSEi	1#	1.89	0.98	1.02	1.02	0.95	1.10	1.12	0.99	1.21	1.69	**1.20**
	2#	2.41	1.14	1.02	0.67	0.99	1.10	0.97	1.03	0.92	0.92	**1.12**

populated in these intervals. (2) The overpopulation of large extreme ranks implies a large under-foresting in the low-frequency interval, especially in the heavy precipitation interval (yellow fan area, 10%). This result indicates that the ensemble precipitation forecast is unreliable. The partial reliability of different precipitation frequency intervals in the ensemble can be verified better using Rpolar diagram by comparison. The values of $RMSE_i$ are 5.62, 3.84, 2.38, 1.37, 1.39, 1.26, 1.23, 1.26, 1.27 and 1.24 for ensemble precipitation forecasts in day 1. The *mRMSE* value is 2.09, which is much larger than 0.70 (Fig. 7b). The larger $RMSE_i$ $(i = 1, 2, 3)$ values at 1st to 3rd quantiles also imply unreliability in the ensemble.

4. Conclusions

The verification of partial reliability of extreme event ensemble forecasts is more meaningful than the overall reliability. Guaranteeing partial reliability in specific intervals has practical significance, such as the evaluation of partial reliability in the high flow situation, low-flow situation, heavy rainfall situation and others. This paper propose an improved verification method using Rpolar diagram.

Previously, uncritical use of rank histograms may result in an illusory understanding of the qualities of the ensemble and is only capable of evaluating the overall reliability of a forecast and is unable to reveal the unreliable forecast performance of extreme events in the ensemble. Nevertheless, the Rpolar diagrams provide an effective solution for verifying the partial reliability in certain frequency intervals, including extremes. Rpolar diagram are useful for determining both overall and partial reliability of ensemble forecasts. Especially when verifying the partial reliability, the shape of the

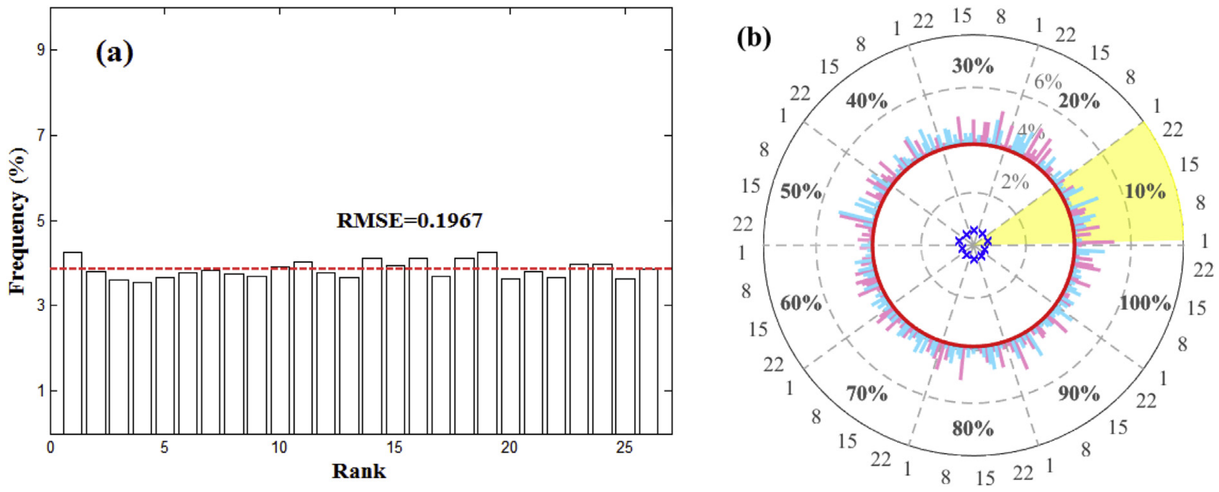

Fig. 6. (a) Rank histogram for perfect reliable ensemble streamflow simulation for the B3 river basin. (b) Corresponding Rpolar diagram.

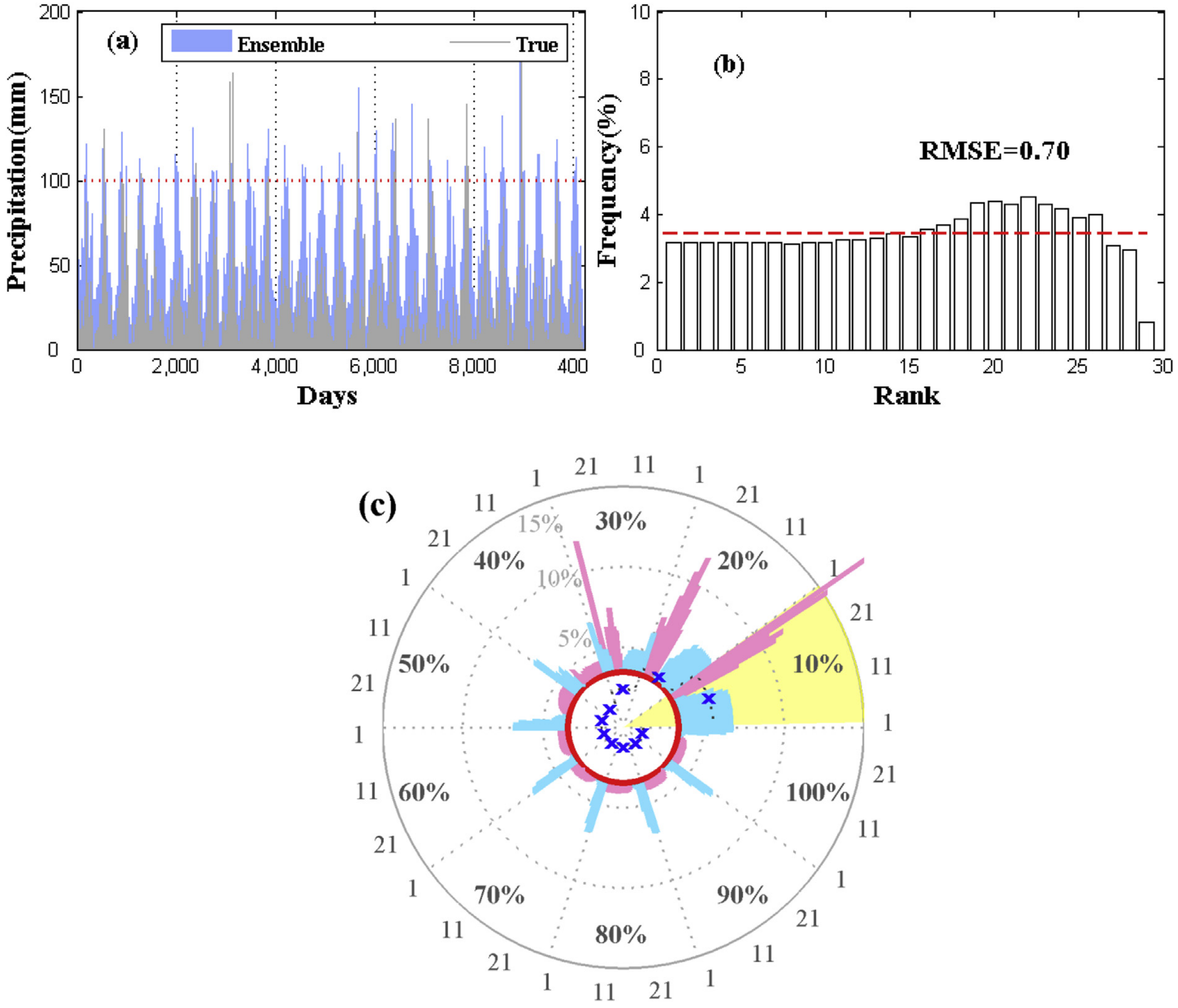

Fig. 7. (a) Schematic of precipitation for the ensemble precipitation forecasts on day 1 (1981–2008). The simulated precipitation data are generated from GFS single-valued forecasts using EPP. Observations are from 187 weather stations in the Huai River Basin in China. (b) Corresponding rank histogram. The horizontal red dashed lines indicate perfect uniformity. (c) Corresponding Rpolar diagram. Perfect uniformity is shown as a red circle. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

polar ranks may provide the users about the characteristics (i.e. over forecasting, under forecasting, no error, ect.) of the ensemble at each quantile. Further applications of Rpolar diagrams in different hydro-climate regimes are also constructed. The MOPEX streamflow simulation daily data and the precipitation daily data preprocessed using EPP are used. The results in both streamflow and precipitation applications show that the Rpolar diagram is useful and effective in verifying the partial reliability of extreme event ensemble forecasts, such as high discharge and heavy precipitation. Note that a special ranks assigning method is also introduced when plotting Rpolar diagrams for the ensemble precipitation forecasts in consideration of the no-precipitation condition.

There is an assumption of “stationarity” about rank histogram and Rpolar diagram. That is to say, the perfect ensemble is that the distribution of the forecast ensemble matches the distribution of historical observations. However, the climate/hydrology is non-stationary as a result of both natural variability and anthropogenic factors (Razavi et al., 2015). Almost all hydrological models have the problem. Rpolar diagram can show the different between ensemble forecast and observations, and then gives some modeling advice (such as increasing the runoff if under-forecasting or decreasing the runoff if over-forecasting). Improved models and reconstructed observation are common methods on non-stationary problem (Toonen, 2015; Mondal and Mujumdar, 2015). Miao et al. (2016) developed an updated non-stationary bias-correction method which combines two widely used quantile mapping methods to eliminate potential illogical values of the variable. The nature discharge is equal to the sum of observed discharge and water use, which is a simple reconstructed observation method. The matter should be considered in further studies.

Acknowledgments

This study was supported by the Natural Science Foundation of China (No. 41475093, 51279210), the State Key Laboratory of Severe Weather Open Research Program (NO. 2015LASW-A05) and the Beijing Science and Technology Plan Project (No. Z141100003614052).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2017.02.024.

References

Abdi, H., 2007. The Kendall Rank Correlation Coefficient. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA, pp. 508–510.
Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Wea. Rev. 78, 1–3.
Beck, O., Repke, D.B., Faull, K.F., 1986. 6-Hydroxymethtryptoline is naturally occurring in mammalian urine: identification by combined chiral capillary gas chromatography and high resolution mass spectrometry. Biol. Mass Spectrom. 13, 469–472.
Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Noise Reduction in Speech Processing, vol. 2. Springer Science & Business Media.
Biondi, D., De Luca, D.L., 2013. Performance assessment of a Bayesian Forecasting System (BFS) 376 for real-time flood forecasting. J. Hydrol. 479, 51–63.
Brown, J.D., Demargne, J., Seo, D.-J., Liu, Y., 2010. The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ. Model. Softw. 25, 854–872.
Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Model. Softw. 22, 1034–1052.
Duan, Q., Schaake, J., Andreassian, V., et al., 2006. Model Parameter Estimation Experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J. Hydrol. 320, 3–17.
Dong, L., Xiong, L., Yu, K., 2013. Uncertainty analysis of multiple hydrologic models using the bayesian model averaging method. J. Appl. Math. 2, 2–22. http://dx.doi.org/10.1155/2013/346045.
Gal, G., Makler-Pick, V., Shachar, N., 2014. Dealing with uncertainty in ecosystem model scenarios: application of the single-model ensemble approach. Environ. Model. Softw. 61, 360–370.
Hamill, T.M., 1997. Reliability diagrams for multicategory probabilistic forecasts. Wea. Forecast. 12, 736–741.
Hamill, T.M., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Mon. Wea. Rev. 129, 550–560.
Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Wea. Forecast. 15, 559–570.
Hoffmann, A., Turelli, M., Harshman, L.G., 1990. Factors affecting the distribution of cytoplasmic incompatibility in Drosophila simulans. Genetics 126, 933–948.
Jolliffe, I.T., Stephenson, D.B., 2008. Proper scores for probability forecasts can never be equitable. Mon. Wea Rev. 136, 1505–1510.
Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth. Syst. Sc. 11, 1267–1277.
Liu, Y., Duan, Q., Zhao, L., Ye, A., Tao, Y., Miao, C., Mu, X., Schaake, J.C., 2013. Evaluating the predictive skill of post-processed NCEP GFS ensemble precipitation forecasts in China's Huai river basin. Hydrol. Process 27, 57–74.
Miao, C.Y., Su, L., Sun, Q.H., Duan, Q.Y., 2016. A nonstationary bias-correction technique to remove bias in GCM simulations. J. Geophys. Res. 121 (10), 5718–5735.
Mondal, A., Mujumdar, P.P., 2015. Modeling non-stationarity in intensity, duration and frequency of extreme rainfall over India. J. Hydrology 521, 217–231.
Murphy, A.H., 1973. A new vector partition of the probability score. J. Appl. Meteorol. Clim. 12, 595–600.
Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Mon. Wea. Rev. 115, 1330–1338.
Mason, S.J., Graham, N.E., 1999. Conditional probabilities, relative operating characteristics, and relative operating levels. Wea. Forecast. 14, 713–725.
Maraun, D., 2013. Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. J. Clim. 26, 2137–2143.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I-A discussion of principles. J. Hydrol. 10, 282–290.
Potts, J.M., Folland, C.K., Jolliffe, I.T., Sexton, D., 1996. Revised “LEPS” scores for assessing climate model simulations and long-range forecasts. J. Clim. 9, 34–53.
Pushpalatha, R., Perrin, C., Moine, N.L., Andreassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. J. Hydrol. 420, 171–182.
Razavi, S., Elshorbagy, A., Wheater, H., Sauchyn, D., 2015. Toward understanding nonstationarity in climate and hydrology through tree ring proxy records. Water Resour. Res. 51, 1813–1830.
Swets, J.A., 1973. The relative operating characteristic in psychology. Science 182, 990–1000.
Sircombe, K.N., 2004. AgeDisplay: an EXCEL workbook to evaluate and display univariate geochronological data using binned frequency histograms and probability density distributions. Comput. Geosci-UK 30, 21–31.
Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., Seo, D.J., 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. Hydrology Earth Syst. Sci. Discuss. Discuss. 4, 655–717.
Stephenson, D.B., Coelho, C.A.S., Jolliffe, I.T., 2008. Two extra components in the Brier score decomposition. Wea. Forecast. 23, 752–757.
Singh, A., Singh, S., Nema, A.K., Singh, G., Gangwar, A., 2014. Rainfall-runoff modeling using MIKE 11 NAM model for vinayakpur intercepted catchment, Chhattisgarh. Indian J. Dryland Agric. Res. Dev. 29, 1–4.
Talagrand, O., Vautard, R., Strauss, B., 1997. Evaluation of probabilistic prediction systems. Proc. ECMWF Workshop Predict. 10, 1–25.
Toonen, W.H.J., 2015. Flood frequency analysis and discussion of non-stationarity of the Lower Rhine flooding regime (AD 1350–2011): using discharge data, water level measurements, and historical records. J. Hydrology 528, 490–502.
Van Steenbergen, N., Ronsyn, J., Willems, P., 2012. A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication. Environ. Model. Softw. 33, 92–105.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79.
Weigel, A.P., Liniger, M.A., Appenzeller, C., 2007. The discrete Brier and ranked probability skill scores. Mon. Wea. Rev. 135, 118–124.
Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences, vol. 100. Academic press.
Yuan, X., Wood, E.F., Roundy, J.K., Pan, M., 2013. CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. J. Clim. 26, 4828–4847.
Ye, A., Duan, Q., Yuan, X., Wood, E., Schaake, J., 2014. Hydrologic post-processing of MOPEX streamflow simulations. J. Hydrol. 508, 147–156.
Zar, J.H., 1998. Spearman rank correlation. Encyclopedia Biostatistics.