Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR032631

A Mutual Information Theory-Based Approach for Assessing Uncertainties in Deterministic Multi-Category Precipitation Forecasts

Yawei Ning[1,2], Guohua Liang[1], Wei Ding[1], Xiaogang Shi[2], Yurui Fan[3], Jianxia Chang[4], Yimin Wang[4], Bin He[1], and Huicheng Zhou[1]

[1]School of Hydraulic Engineering, Dalian University of Technology, Dalian, China, [2]School of Interdisciplinary Studies, University of Glasgow, Dumfries, UK, [3]Department of Civil and Environmental Engineering, Brunel University, London, UK, [4]State Key Laboratory of Eco-hydraulics in Northwest Arid Region of China, Xi'an University of Technology, Xi'an, China

Key Points:

- A mutual information theory-based approach was developed for assessing uncertainties in deterministic multi-category precipitation forecasts
- The proposed approach shows a better performance than some traditional verification methods
- The proposed approach needs a careful choice of bin width

Correspondence to:

X. Shi,
John.Shi@glasgow.ac.uk

Citation:

Ning, Y., Liang, G., Ding, W., Shi, X., Fan, Y., Chang, J., et al. (2022). A mutual information theory-based approach for assessing uncertainties in deterministic multi-category precipitation forecasts. *Water Resources Research*, *58*, e2022WR032631. https://doi.org/10.1029/2022WR032631

Received 20 APR 2022
Accepted 14 NOV 2022

Author Contributions:

Conceptualization: Yawei Ning, Xiaogang Shi
Data curation: Yawei Ning
Investigation: Yawei Ning
Resources: Guohua Liang, Xiaogang Shi, Bin He, Huicheng Zhou
Supervision: Guohua Liang, Wei Ding, Xiaogang Shi, Bin He, Huicheng Zhou
Writing – original draft: Yawei Ning
Writing – review & editing: Wei Ding, Xiaogang Shi, Yurui Fan, Jianxia Chang, Yimin Wang

Abstract The very nature of weather forecasts and verifications and the way they are used make it impossible for one single or absolute standard of evaluation. However, little research has been conducted on verifying deterministic multi-category forecasts, which is based on the attribute of uncertainty. The authors propose a new approach using two mutual information theory-based scores for assessing the comprehensive uncertainty of all categories and the uncertainty for a certain category in deterministic multi-category precipitation forecasts, respectively. Specifically, the comprehensive uncertainty is defined as the average reduction in uncertainty about the observations resulting from the use of a predictive model to provide all categories forecasts; the uncertainty of a certain category is defined as the reduction in uncertainty about the observations resulting from the use of a predictive model to provide a certain category forecast. By applying the proposed approach and traditional verification methods, the four precipitation forecasting products from the China Meteorological Administration, European Centre for Medium-Range Weather Forecasts, National Centers for Environmental Prediction, and United Kingdom Meteorological Office were verified in the Dahuofang Reservoir Drainage Basin, China. The results indicate that: (a) the proposed approach can better capture the changing patterns of uncertainties with lead times and distinguish the forecasting performance among different forecast products; (b) the proposed approach is resistant to the extreme bias; (c) the proposed approach needs a careful choice of bin width; and (d) the bias analysis is necessary before verifying the uncertainties in precipitation forecasts.

1. Introduction

With the development of numerical weather prediction (NWP), various types of precipitation forecasts have been developed and the quality of precipitation forecasting has been continuously improved (Dance et al., 2019; North et al., 2013; Rodwell et al., 2010; Sharma et al., 2021). Generally, precipitation forecasts can be classified into deterministic forecasts, probabilistic forecasts and ensemble forecasts (Jolliffe & Stephenson, 2012, pp. 11–12; Shi et al., 2008; Xu et al., 2020). It should be noted that ensemble forecasts can also be regarded as probabilistic forecasts that are expressed as a discrete approximation to a full forecast probability density function (Wilks, 2019, p. 433). In terms of data formats, they can also be classified into continuous forecasts and categorical (binary or multi-category) forecasts (Jolliffe & Stephenson, 2012, p. 11; Murphy & Winkler, 1987). Probabilistic forecasts, which enhance useful uncertainty information, became available to the public in the United States for more than 50 years (Murphy, 1998). Nevertheless, probabilistic forecasts are still difficult to be correctly understood by the public (Davis et al., 2016, p. 95; Fundel et al., 2019; Ishikawa et al., 2005; Joslyn et al., 2009; Murphy, 1998). By contrast, deterministic forecasts are simpler and easier to use for end-users (Fundel et al., 2019). In particular, deterministic multi-category precipitation forecasts are commonly used in water resources management and other related fields, such as avalanche and flood warnings (Economou et al., 2016; Schirmer & Jamieson, 2015), reservoir operation (B. Wang & Zhou, 2006, p. 97; Cai et al., 2019; Ning et al., 2021; Peng et al., 2017; Xi et al., 2010; Zhou et al., 2011) and drought management (Łabędzki, 2017; Sigaroodi et al., 2014). However, the precipitation forecasts from the NWP are biased and uncertain in general, which may bring errors. Therefore, it is crucial to verify deterministic multi-category precipitation forecasts,

which can serve for administrative, scientific, and economic purposes (Brier & Allen, 1951, pp. 841–842; Wilks, 2019, p. 369).

The very nature of weather forecasts and verifications and the way they are used make it impossible for one single or absolute standard of evaluation (Brier & Allen, 1951, p. 843; Brooks & Doswell, 1996; Mason & Weigel, 2009). In previous studies, many techniques have been developed and applied in the verification of forecasts from the attributes of bias, association, accuracy, skill, reliability, resolution, sharpness, discrimination, and uncertainty (Bradley et al., 2016, pp. 7–9; Murphy, 1993). Proper verification methods for deterministic multi-category forecasts are mainly based on contingency tables and include the proportion correct (PC), bias ratio (BR), probability of detection (POD) or hit rate and skill scores (Jolliffe & Stephenson, 2012, p. 64), which are based on the attributes of accuracy, bias, discrimination and skill, respectively (Wilks, 2019, pp. 376–381). However, little research has been conducted on verifying deterministic multi-category forecasts, which is based on the attribute of uncertainty. Using the interquartile range (IQR), Brown and Murphy (1987) verified the fire-weather forecasts from the attribute of uncertainty, which is defined as variability in the conditional distributions of observed values given each distinct forecast value. Besides the IQR, the standard deviation (Std) are commonly used to measure the variability of distributions (Gong et al., 2013; Jolliffe & Stephenson, 2012, p. 19). However, the Std is neither robust nor resistant (Wilks, 2019, p. 27). Even one very large value would affect the Std very much because it is especially far away from the mean and the difference can be further magnified by the squaring process (Wilks, 2019, p. 27). The IQR is very easy to compute, but it has the disadvantage of not making much use of a substantial fraction of the data (Wilks, 2019, p. 27). In recent years, however, there has been no further development on verification methods for deterministic multi-category forecasts (Casati et al., 2008; Dorninger et al., 2020; Ebert et al., 2013; Gilleland et al., 2016, pp. 6–8; Wilks, 2019, pp. 388–394). Therefore, more resistant and accurate verification methods are needed to measure the uncertainties in deterministic multi-category forecasts.

Information entropy is a natural and fundamental measure of uncertainty in a number of fields, including water engineering (Singh, 2013, p. 310). The information entropy is calculated by the frequency (or probability) distributions of the analyzed samples rather than directly calculated by the values of analyzed samples (like Std) or partly quantiles of the analyzed samples (like IQR). Therefore, compared with Std and IQR, information entropy is less sensitive to extreme values than Std, and can utilize more information from the analyzed samples than IQR. Further, information entropy can provide a more accurate characterization of uncertainty than Std, since the latter depends only on the second moment, whereas information entropy takes into account the effects of higher order moments (Gong et al., 2013). DelSole and Tippett (2007) pointed out that it is difficult to conceive of a measure better suitable for a general evaluation of uncertainty than information entropy. Mutual information, an entropy-based statistic, characterizes what one (uncertain) variable X can tell us about another (uncertain) variable Y; in other words, how much information is shared between the two, or how much of the uncertainty about Y can be reduced by knowing X (Gong et al., 2013). Mutual information has been used in the verification of probabilistic weather forecasts, such as the ranked mutual information scores (Ahrens & Walser, 2008) and the resolution component of divergence score (DS), which is the mutual information between forecasts and observations (Weijs, Schoups, et al., 2010; Weijs, van Nooijen, et al., 2010). Both the ranked mutual information scores and DS are metrics for the verification of probabilistic forecasts. However, they are not applied for deterministic forecasts. In addition, these previous studies have neglected the proper choice of bin width when calculating mutual information, which would bring extra errors because entropy estimates with inappropriate bin width are positively biased (Pechlivanidis et al., 2016; Ruddell & Kumar, 2009).

This paper aims to propose a new approach using two mutual information theory-based scores to assess the uncertainties in deterministic multi-category precipitation forecasts, that can cover both the comprehensive uncertainty of all categories and the uncertainty of a certain category. The comprehensive uncertainty is defined as the average reduction in uncertainty about the observations resulting from the use of a predictive model to provide all categories forecasts; the uncertainty of a certain category is defined as the reduction in uncertainty about the observations resulting from the use of a predictive model to provide a certain category forecast. It should be noted that the two scores are for the verification purpose of forecasting products. The Dahuofang Reservoir was used as an example, as shown in Section 2, for the verification of daily precipitation forecasts. Section 3 presents the new scores and other reference scores. Section 4 shows the verification results. The discussion and conclusions are summarized in Sections 5 and 6, respectively.

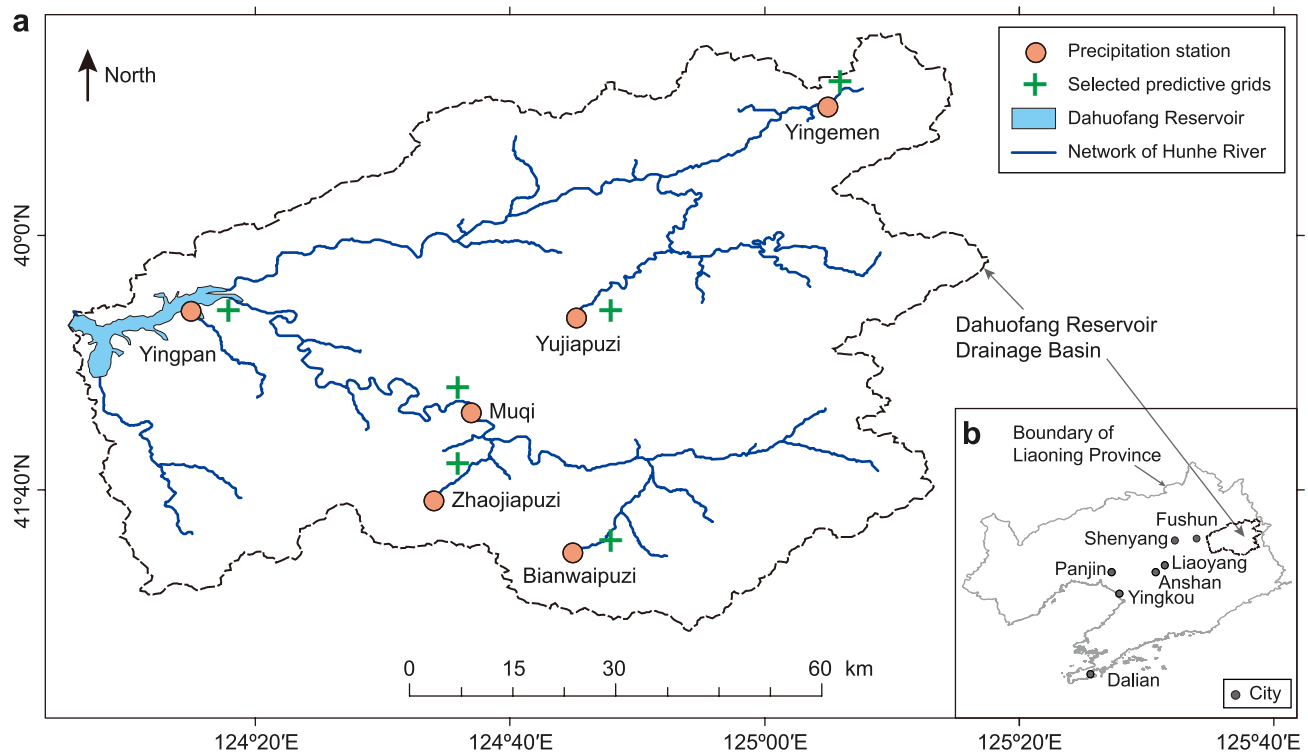

Figure 1. Schematic of (a) the Dahuofang Reservoir Drainage Basin with the location of precipitation stations and (b) Liaoning Province.

2. Study Area and Data

2.1. Study Area

The Dahuofang Reservoir Drainage Basin (DRDB), with a drainage area of 5,437 km^2, is located in Liaoning Province, Northeast China, as shown in Figure 1. The Reservoir not only undertakes the flood control of protecting Shenyang city and Fushun city but also provides water to the cities of Shenyang, Fushun, Liaoyang, Anshan, Panjin, Yingkou, and Dalian (shown in Figure 1b). Six precipitation stations (shown in Figure 1a), namely Bianwaipuzi, Muqi, Yingemen, Yingpan, Yujiapuzi, and Zhaojiapuzi, were selected in the study. As shown in Figure 2a, the annual cycle of monthly precipitation at the six precipitation stations is consistent except for some differences in the summer months. Therefore, the spatial variability of precipitation across the DRDB is not significant. The average annual precipitation in the DRDB is 812 mm which is unevenly distributed during the year. As shown in Figure 2b, almost 85% of annual precipitation at the Muqi precipitation station (shown in Figure 1a) falls in the flooding season from May to October for the period 2007–2018, which resulted in most of the floods in the study area. It should be noted that the Muqi precipitation station in Figure 2b was randomly selected as one example.

2.2. Forecast and Observation Data

The Observing System Research and Predictability Experiment Interactive Grand Global Ensemble (TIGGE) data set was used in this study, which is a global NWP data set and has been widely used in the scientific research because of its comprehensiveness and public availability (Swinbank et al., 2016). The control forecasts (single deterministic forecasts) of precipitation from the four data products were selected in this study, including the China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP), and United Kingdom Meteorological Office (UKMO). More details of these products used in this paper are briefly given in Table 1.

All the observed precipitation data derived from the six precipitation stations (shown in Figure 1a) in the study were obtained from the Hydrological Yearbook of the People's Republic of China. To be consistent with the

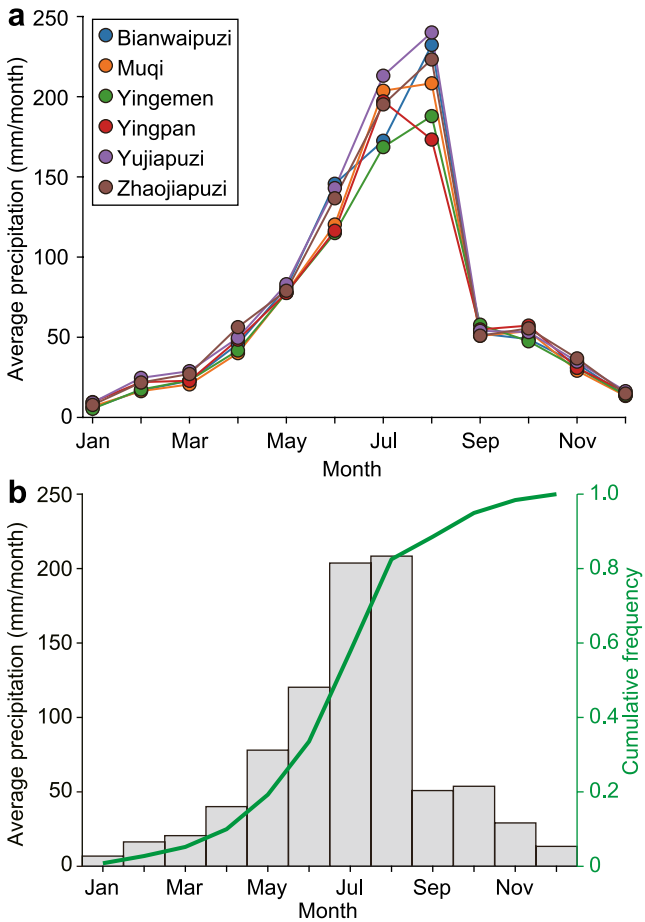

Figure 2. Monthly precipitation over the period of 2007–2018 in Dahuofang Reservoir Drainage Basin. (a) The Monthly precipitation from the six precipitation stations; and (b) the monthly precipitation and cumulative frequency estimates from the precipitation station at Muqi.

observed precipitation, we only selected the forecasts on the base time of 00:00 UTC. The range of lead time for forecast precipitation used in the study is from 1 to 7 days. As the horizontal resolution is different for the above four forecast products, it is inconvenient to compare them with the observed precipitation. Therefore, the original precipitation forecast data of CMA, ECMWF, NCEP, and UKMO are converted to a 0.1° × 0.1° grid using the bilinear interpolation software provided by the ECMWF TIGGE data portal (Su et al., 2014). The interpolated data grid, the center of which is closest to a specific precipitation station, is used as the source of forecast precipitation (H. Wang et al., 2021).

The study focuses on the flooding season from May to October over the period 2007–2018 with 2,208 days in total. At each precipitation station, there is no missing data and the observed sample size is 2,208. However, the number of missing values in the forecasts differs among different products and lead times. The detailed sample sizes are shown in Table 2 for different products and lead times.

2.3. Classification of Forecasts of Daily Precipitation

The objective of the study is to assess uncertainties in deterministic multi-category forecasts. However, the precipitation forecasts from the TIGGE data set are continuous and need to be classified. The classification of precipitation forecasts differs with its applications. In this study, we used the classification of deterministic multi-category precipitation forecasts, which are applied in reservoir operation as an example. Table 3 indicates the CMA classification standards for daily precipitation. Using the CMA classification standards, we can obtain the total number of samples for each category from observed precipitation from 2007 to 2018 (May to October), as shown in Table 4. It can be seen from Table 4 that the sample number is very small for heavy rain and higher categories. We combined the precipitation data greater than 25 mm into category L3 (≥25.0 mm). The events with no rain and light rain were combined into category L1 (0–9.9 mm) in this study because these two categories are often used together when applied in reservoir operation (B. Wang & Zhou, 2006, p. 98; Cai et al., 2019; Peng et al., 2017; Xi et al., 2010; Zhou et al., 2011). Finally, the events with medium rain were classified to category L2 (10.0–24.9 mm). The new classification is shown in Table 5.

3. Methods

By using mutual information theory-based scores, a new approach was proposed to measure the comprehensive uncertainty of all categories and the uncertainty for a certain category, as shown in Section 3.1. We compared the proposed approach with several traditional methods in Section 3.2. To provide a better understanding for replication, we made one example to clearly show each step for using the proposed method. The access to the example data and MATLAB code is shown in Data Availability Statement.

Table 1
Characteristics of the Four Forecast Products Used in This Study

Products	Time range (days)	Resolution (°)	Base time (UTC)
CMA	0–15	0.2815 × 0.2812	0/12
ECMWF	0–15	O640 (ORGG)	0/12
NCEP	0–16	1 × 1	0/6/12/18
UKMO	0–7.25	0.187 × 0.2815	0/6/12/18

Note. ORGG: octahedral reduced Gaussian grid.

3.1. A Mutual Information Theory-Based Approach

3.1.1. The Assessment of Comprehensive Uncertainty for All Categories

The normalized mutual information (NMI) for two variables, also called the "uncertainty coefficient" (Press et al., 1986), can be interpreted as the relative reduction in uncertainty about one variable from "getting to know another"

Table 2
Sample Size of Different Products and Lead Times From 2007 to 2018 (May to October)

	Lead times (day)						
Products	1	2	3	4	5	6	7
CMA	2,083	1,989	1,998	2,000	2,017	1,994	1,994
ECMWF	2,208	2,085	2,085	2,085	2,085	2,085	2,085
NCEP	2,177	2,177	2,037	2,037	2,148	2,061	2,090
UKMO	2,100	1,946	1,946	1,977	1,977	1,977	1,824

(Särndal, 1974; Topp et al., 2013). Thus, the NMI for observed precipitation and forecast precipitation is used to measure the comprehensive uncertainty of all categories and is defined in Equation 1

$$\mathrm{NMI} = \frac{H(O) - H(O|F)}{H(O)} = \frac{I(O;F)}{H(O)} \tag{1}$$

where O and F represent observed precipitation and corresponding forecast precipitation, respectively; $H(O)$ is the entropy of O that represents the uncertainty in O; $H(O|F)$ is the conditional entropy of O given F that represents the amount of uncertainty remaining in O after F is known; $I(O;F)$ is the mutual information of O and F, which represents the amount of uncertainty eliminated in O through observing F. Therefore, NMI represents a ratio of uncertainty eliminated about O resulting from the use of forecast precipitation F (Topp et al., 2013).

When comparing the performance of forecasts under different climatologies, it is more reasonable to consider the degree of difficulty of forecasting (the uncertainties in the observed precipitation) than to not consider it. It should be noted that the word "climatology" refers to the empirical distribution functions of the observable precipitation based on a sample of past observations, that is, relative frequencies of past events (Jolliffe & Stephenson, 2012, p. 247). Thus, NMI is more appropriate than the mutual information for comparing the performance of forecasts of different climatologies, such as forecasts in different areas, because the relative value (NMI), rather than the absolute value (mutual information) can consider the effects of different climatologies.

3.1.2. The Assessment of Uncertainty for a Certain Category

The mutual information can be decomposed by Equation 2 (Hughes et al., 2017; Topp et al., 2013).

$$I(O;F) = \sum_{k=1}^{K} \hat{p}_k \times [H(O) - H(O|F_k)] \tag{2}$$

In Equation 2, K is the number of categories of forecasts and equal to 3 in the study. k is the index of a forecast category and represents category L1, L2, and L3 if k takes a value 1, 2, and 3, respectively. F_k is the kth category forecast precipitation. $\hat{p}_k$ is the frequency of the occurrence of F_k. $O|F_k$ is the observed precipitation O given F_k. $H(O|F_k)$ is the entropy of $O|F_k$ and represents the amount of uncertainty remaining about O, after receiving the kth category precipitation forecast information.

With Equation 2, we can decompose NMI in Equation 1 as follows

$$\mathrm{NMI} = \sum_{k=1}^{K} \hat{p}_k \times \mathrm{NMI}_k \tag{3}$$

$$\mathrm{NMI}_k = \frac{H(O) - H(O|F_k)}{H(O)} \tag{4}$$

where NMI_k represents a ratio of uncertainty eliminated about O after receiving the kth category precipitation forecast information and is used to measure the uncertainty of a certain category, namely the kth category. The larger NMI_k, the smaller uncertainty forecasts in the kth category.

Table 3
Classification Standard of Daily Precipitation by China Meteorological Administration (CMA)

Magnitude	Classification standard of precipitation	Amount of daily precipitation (mm)
1	No rain	0–0.1
2	Light rain	0.1–9.9
3	Medium rain	10.0–24.9
4	Heavy rain	25.0–49.9
5	Rainstorm	50.0–99.9
6	Heavy rainstorm	100.0–249.9
7	Extreme rainstorm	⩾250.0

3.1.3. The Calculation Procedures of Two Mutual Information Theory-Based Scores

The calculation of two mutual information theory-based scores at a certain precipitation station contains four steps, as shown in Figure 3. The values of observed precipitation and the categories of forecast precipitation are pseudo which are taken just for an example.

3.1.3.1. Step 1: Calculate the Bin Width

The calculation of entropy requires a careful choice of bin width (Gong et al., 2014) and the methods to properly estimate the bin width include

Table 4
The Number of Samples of Observed Precipitation From 2007 to 2018 (May to October) at the Six Precipitation Stations

Precipitation station	No rain	Light rain	Medium rain	Heavy rain	Rainstorm	Heavy rainstorm	Extreme rainstorm	Sum of samples
Bianwaipuzi	1,414	552	156	70	14	2	0	2,208
Muqi	1,456	522	149	56	22	3	0	2,208
Yingemen	1,412	568	163	48	16	1	0	2,208
Yingpan	1,444	541	147	57	15	4	0	2,208
Yujiapuzi	1,383	560	181	64	15	4	1	2,208
Zhaojiapuzi	1,418	552	144	72	19	3	0	2,208

binning with fixed mass, fixed width or hybrid fixed width–mass interval partitions (Pechlivanidis et al., 2016). Fixed width bins have the advantage of being simple and computationally efficient (Pechlivanidis et al., 2016; Ruddell & Kumar, 2009; Thiesen et al., 2019, 2020). In this study, we introduce five fixed width binning methods, as shown in Table 6.

In Table 6, the relationship between W and NC is shown in Equation 5

$$W = \frac{R}{\mathrm{NC}} \tag{5}$$

where R is the range of distribution.

When calculating mutual information (or NMI) of two variables, there is often only one binning strategy for any variable (Alfonso et al., 2010; Babel et al., 2015; Mogheir et al., 2003). In the study, because $O|F_k$ is a part of O, the bin width to calculate $H(O)$ and $H(O|F_k)$ should be the same. Therefore, the data of O at each station which contain all observed precipitation are used to calculate the bin width. With the bin width, the range of each bin can be obtained. It should be noted that the bin width calculated by fixed width binning methods may differ among different precipitation stations. In addition to the fixed width binning method, the authors also studied one non-fixed binning method (denoted as M0). The binning strategy of M0 in the study is the same as the classification of forecasts. Therefore, the binning method M0 contains three bins, namely L1 (0–9.9 mm), L2 (10.0–24.9 mm), and L3 (≥25.0 mm), as summarized in Table 5.

3.1.3.2. Step 2: Bin Each Observation and Classify Each Forecast

With the bin width, observed precipitation can be put into individual bins. Take the case in Figure 3 as an example. If the bin width calculated is 3 mm, the intervals of the bins are 0–2.9 mm (C_1), 3–5.9 mm (C_2), 6–8.9 mm (C_3), 9–11.9 mm (C_4), 12–14.9 mm (C_5), 15–17.9 mm (C_6), 18–20.9 mm (C_7)…. Therefore, the observed precipitation {0, 20, 11, 2} mm are put into bins {C_1, C_7, C_4, C_1}, respectively. Similarly, the forecasts can be classified into different categories with the classification of forecasts. The observations and corresponding forecasts each day are put into one certain bin and category, respectively, as shown in Figure 3.

Table 5
Classification of the Forecasts of Daily Precipitation Used in the Study

Classification standard of precipitation	Amount of daily precipitation (mm)
L1	0–9.9
L2	10.0–24.9
L3	⩾25.0

3.1.3.3. Step 3: Calculate the Normalized Contingency Table

With the binning results and classification of forecasts, we can calculate $n_{k,j}$ (k = L1, L2, L3; j = C_1, C_2, C_3, …, C_{NC}), which represents the number of the samples (days) when forecasts and observed precipitation belong to the kth category and the jth class interval, respectively. Thus, the contingency table can be obtained, which gives the discrete joint sample distribution of deterministic forecasts and categorical observations in terms of cell counts (Jolliffe & Stephenson, 2012, pp. 62–63, 243). Furthermore, $p_{k,j}$ can be obtained by Equation 6

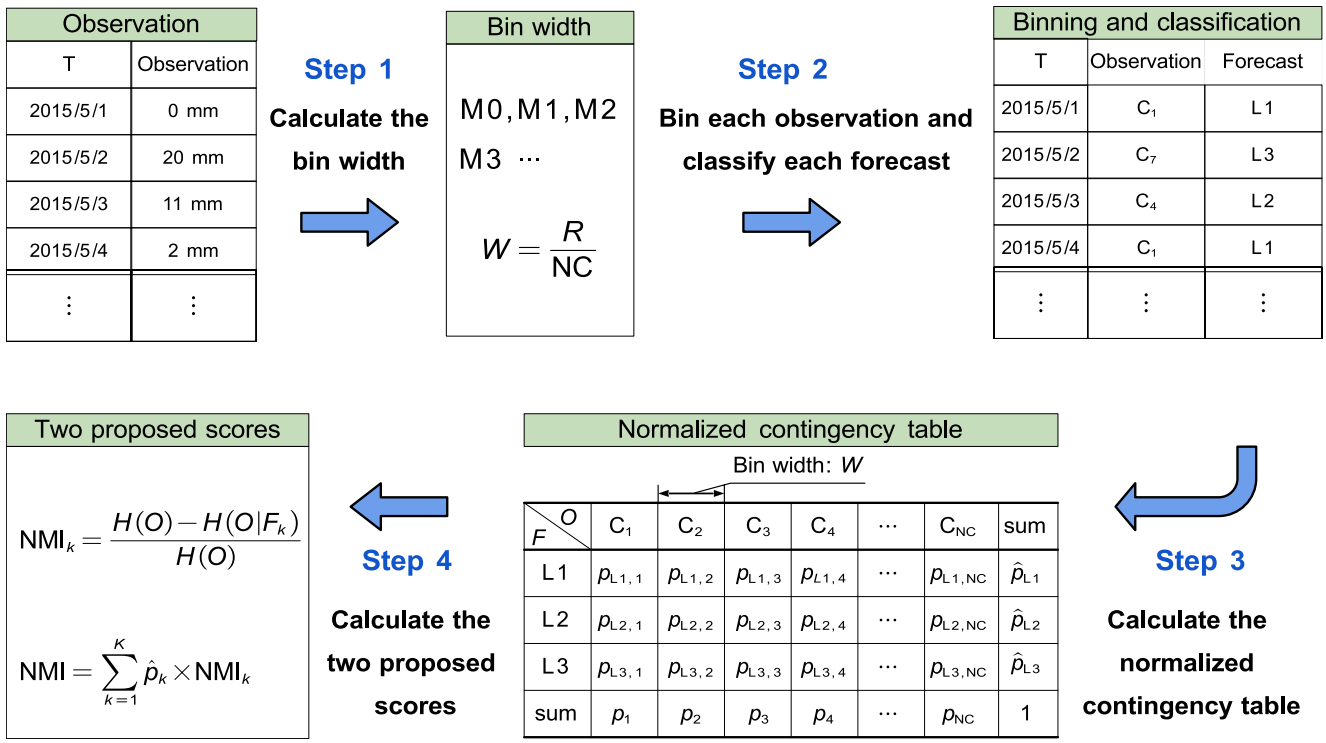

Figure 3. The schematic of the procedure to calculate the two proposed scores.

$$p_{k,j} = \frac{n_{k,j}}{N} \tag{6}$$

where N denotes the sample size; $p_{k,j}$ is the joint frequency when the forecasted precipitation and observed precipitation belong to kth category and jth class interval, respectively. With $p_{k,j}$, the normalized contingency table can be obtained, as shown in Table 7.

3.1.3.4. Step 4: Calculate the Two Proposed Scores

In Equations 3 and 4, NMI_k and NMI can be obtained after $\hat{p}_k$, $H(O)$ and $H(O|F_k)$ are known. $\hat{p}_k$ can be estimated by the proportion F_k in all forecasts. Based on a normalized contingency table (Table 7), $H(O)$ and $H(O|F_k)$ can be obtained by Equations 7 and 8, respectively.

Table 6
Five Fixed Width Binning Methods to Calculate Entropy

Binning methods ID	Binning methods	Related studies
M1	$W = 3.49 \times \sigma \times S^{-\frac{1}{3}}$	Gong et al. (2014), Loritz et al. (2019), Scott (1979), and Singh (2013, p. 544)
M2	$NC = Int(1 + \log_2 S)$	Sturges (1926) and W. Wang et al. (2018)
M3	$NC = Int(1 + 1.33 \times \log_2 S)$	Masoumi and Kerachian (2010), Mogheir et al. (2003), and Ridolfi et al. (2011, 2016)
M4	$NC = Int(\sqrt{S})$	Hao and Singh (2013) and Montgomery and Runger (2014, p. 135)
M5	The method based on maximum a posteriori estimation and Bayes' theorem (also called the Generalized Knuth method)	Gencaga et al. (2015), Knuth (2013), and Tovo et al. (2016)

Note. W: the bin width; NC: the number of classes (bins); σ: the standard deviation of the distribution; S: the number of available samples belonging to the distribution; Int(·): the decimal integer function; The MATLAB code of M5 can be found in Knuth (2013).

Table 7
Schematic of Normalized Contingency Table in the Study

	Observed						
Forecast	C_1	C_2	C_3	C_4	⋯	C_{NC}	$\sum$
$L1$	$p_{L1,1}$	$p_{L1,2}$	$p_{L1,3}$	$p_{L1,4}$	⋯	$p_{L1,NC}$	$\hat{p}_{L1}$
$L2$	$p_{L2,1}$	$p_{L2,2}$	$p_{L2,3}$	$p_{L2,4}$	⋯	$p_{L2,NC}$	$\hat{p}_{L2}$
$L3$	$p_{L3,1}$	$p_{L3,2}$	$p_{L3,3}$	$p_{L3,4}$	⋯	$p_{L3,NC}$	$\hat{p}_{L3}$
$\sum$	p_1	p_2	p_3	p_4	⋯	p_{NC}	1

Note. C_j: the jth class interval of observed precipitation; p_j: the climatology probability when observed precipitation of the jth class interval occurs; $p_{k,j}$: the joint frequency when forecasts and observed precipitation belong to the kth category and the jth class interval, respectively.

$$H(O) = -\sum_{j=1}^{NC} p_j \log_2 p_j \tag{7}$$

$$H(O|F_k) = -\sum_{j=1}^{NC} p_{j|F_k} \log_2 p_{j|F_k} \tag{8}$$

The logarithm of base two was chosen so that the entropy is expressed in bits; the meaning of p_j is shown in Table 7; $p_{j|F_k}$ is the conditional frequency when observed precipitation belongs to the jth class interval given F_k and can be calculated by Equation 9

$$p_{j|F_k} = \frac{p_{k,j}}{\hat{p}_k} \tag{9}$$

3.1.4. The Values of the Proposed Scores for Optimal Forecasts

In the study, the optimal forecasts denote the forecasts whose values are equal to the values of corresponding observations.

3.1.4.1. The Value of the NMI_k for Optimal Forecasts

NMI_k may be positive ($H(O) > H(O|F_k)$), in which case uncertainty has decreased, or negative ($H(O) < H(O|F_k)$), in which case uncertainty has increased (Hughes & McRoberts, 2014; Hughes et al., 2017; Topp et al., 2013). NMI_k is equal to 0 when the forecasts in kth category are independent with observed precipitation (DelSole & Tippett, 2007). From Equation 3, we found that at least one of the three variables NMI_{L1}, NMI_{L2}, NMI_{L3} is non-negative, because the NMI and $\hat{p}_k$ are non-negative. When forecasts in kth category are optimal, the distribution of $O|F_k$ is the same as O_k, which are the observed precipitation in kth category. Therefore, the NMI_k for optimal forecasts, namely $\mathrm{NMI}_k^{\mathrm{opt}}$, can be calculated as follows

$$\mathrm{NMI}_k^{\mathrm{opt}} = \frac{H(O) - H(O_k)}{H(O)} \tag{10}$$

where $H(O_k)$ is the entropy of O_k.

3.1.4.2. The value of the NMI for optimal forecasts

In Equation (1), $H(O)$, $H(O|F)$, $I(O;F)$ and NMI are all non-negative values. NMI has a value of 0 when the forecasts and observations are independent (Hughes et al., 2017).

Based on Equations 3 and 10, the value of NMI for optimal forecasts, namely $\mathrm{NMI}^{\mathrm{opt}}$, can be calculated by Equation 11

$$\mathrm{NMI}^{\mathrm{opt}} = \sum_{k=1}^{K} \hat{p}_k \times \mathrm{NMI}_k^{\mathrm{opt}} \tag{11}$$

From Equations 7, 8, 10 and 11, we can find that the binning method also affects the values of $\mathrm{NMI}_k^{\mathrm{opt}}$ and $\mathrm{NMI}^{\mathrm{opt}}$, which are no more than 1. $\mathrm{NMI}_k^{\mathrm{opt}}$ is equal to 1 when the observed precipitation in category k are completely within a certain bin while the $\mathrm{NMI}^{\mathrm{opt}}$ is equal to 1 only if the observed precipitation in each forecast category are completely within a different bin. For example, the binning strategy M0 is adopted.

3.2. Comparison of the Proposed Scores and Traditional Scores

3.2.1. Traditional Verification Methods

3.2.1.1. To Assess the Comprehensive Uncertainty of All Categories

Jolliffe and Stephenson (2012, p. 73) pointed out that the GS (the Gerrity score) is an appropriate choice for most ordinal categories event forecast verification problems, which has many desirable properties and is used as a reference verification method for NMI in this study. The GS can be calculated by Equation 12

$$\mathrm{GS} = \sum_{k=1}^{K} \sum_{j=1}^{K} p(k, j) \times s(k, j) \tag{12}$$

where j is the index of an observed category; $p(k, j)$ and $s(k, j)$ are the frequency and the corresponding element of scoring matrix when the forecasts belong to the kth category and observed precipitation belong to the jth category, respectively. The details to calculate the scoring matrix of GS can be seen in Jolliffe and Stephenson (2012, pp. 69–71).

3.2.1.2. To Assess the Uncertainty for a Certain Category

The Std (or variance) and 50% IQR (Wilks, 2019, p. 27), that is, the range between the 25th and the 75th percentiles of the samples, are commonly used to measure uncertainty (Gong et al., 2013) and used as reference verification methods for NMI_k in this study. In particular, the IQR and Std for a certain category can be calculated by Equations 13 and 14, respectively.

$$\mathrm{IQR}_k = f_k^u - f_k^l \tag{13}$$

$$\mathrm{Std}_k = \sqrt{\frac{1}{m-1} \sum_{j=1}^{m} \left(O_k^j - \overline{O}_k\right)^2} \tag{14}$$

In Equations 13 and 14, the IQR_k and Std_k indicate that the 50% IQR and Std of observed precipitation given F_k, respectively; f_k^u and f_k^l are the 75th and 25th percentiles of observed precipitation given F_k; O_k^j denotes the observed precipitation given F_k; m and $\overline{O}_k$ are the number and mean of observed precipitation given F_k, respectively.

3.2.2. Comparison of the Proposed Scores and Traditional Scores

The proposed score NMI and the traditional score GS are used to assess the comprehensive uncertainty of all categories. Both the NMI and GS are calculated by the normalized contingency table, though the corresponding normalized contingency tables are different. Therefore, the two scores are not susceptible to extreme values. The scoring matrix of GS in this study is determined by the daily climatological precipitation frequency in each category (Jolliffe & Stephenson, 2012, pp. 69–71). Therefore, same with NMI, GS is also suitable for comparing the forecasting performance under different climatologies.

The proposed score NMI_k and the traditional score IQR_k and Std_k are used to assess the uncertainty in a certain category. The IQR_k is only affected by the samples at the 25th and 75th percentiles and cannot fully represent the uncertainty of the samples. The Std_k is susceptible to extreme values. Unlike the IQR_k and Std_k, the NMI_k is calculated by the entropy of the relative distribution of forecasting and the distribution of observed precipitation (as shown in Equation 4). Thus, the NMI_k makes full use of the information of all the samples and is less impacted by the extreme values. In addition, the NMI_k can take account of both the uncertainties in the observed precipitation and the eliminated uncertainties (or remaining uncertainties) given the forecasts while both IQR_k and Std_k only calculate the latter. Therefore, the NMI_k is more suitable for comparing the forecasting performance under different climatologies than the IQR_k and Std_k.

From Equation 3, we can find that the NMI is the weighted average value of NMI_k among all categories if we take $\hat{p}_k$ as the weight of NMI_k. Therefore, the score NMI for assessing the comprehensive uncertainty also represents the weighted average uncertainty among all categories. Thus, the new mutual information theory-based approach, as shown in Equations 1, 3 and 4, could verify both the comprehensive performance of all categories of forecast

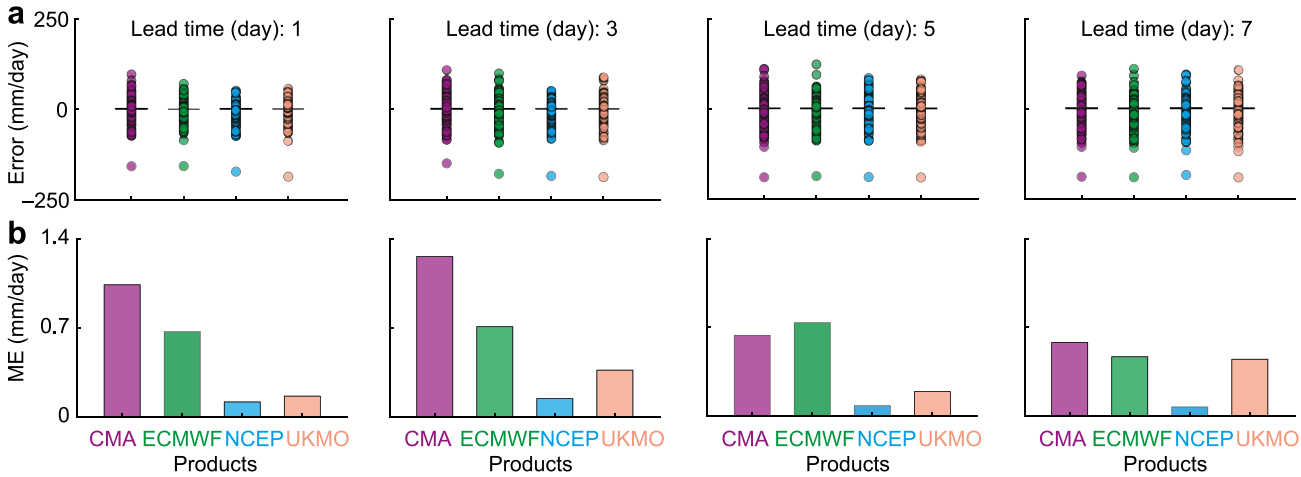

Figure 4. The bias analyses for the four forecast products with the lead times of +1, +3, +5, and +7 days at Muqi: (a) the box and whisker plots of forecast errors; and (b) the mean error of forecasts.

and the forecast performance for a certain category and establish the linkage between these two parts in deterministic multi-category forecasts. However, it is hard for traditional scores to establish the linkage.

4. Results

In this section, we compared the proposed method with traditional verification methods using three criteria:

1. First, a good verification method should clearly distinguish the forecasting performance among different forecast products because selecting the higher quality product or rejecting the lower quality product is important for end-users.
2. Second, a suitable verification method should be stable. Thus, it should not fluctuate much with lead times and can well capture the changing patterns of uncertainties with lead times.
3. Third, a good score should be resistant to extreme bias because a small number of extreme values can have undue influence on the values of scores, and mask the quality of forecasts for non-extreme values (Jolliffe & Stephenson, 2012, p. 7).

It should be pointed out that these three criteria are related. For example, it is difficult for verification methods with poor stability or poor resistivity to clearly distinguish the forecasting performance among different forecast products and well capture the changing patterns of uncertainties with lead times.

4.1. Extreme Bias in the Forecasts

The box and whisker plots of forecast errors with the lead times of +1, +3, +5, and +7 days at Muqi are presented in Figure 4a. Figure 4b shows the corresponding mean error (ME). It should be noted that the monthly precipitation and the forecasting performance at the six precipitations are similar. Therefore, the Muqi precipitation station was randomly selected as one example. The ME values in Figure 4b are positive but no more than 1.4 mm. Therefore, the precipitation from the four forecasting products is not significantly larger than the observed precipitation. However, we can find that a few outliers, that is, extreme bias, exist in Figure 4a. Therefore, the extreme bias cannot be ignored and should be taken into consideration in the following analysis.

4.2. Verification for Comprehensive Uncertainty of All Categories

The proposed score NMI and the GS were used to assess the comprehensive uncertainty of all categories. Figure 5 shows the NMI and GS of four forecast products at the six precipitation stations. The NMI and GS show similar results: (a) ECMWF performs best and NCEP performs worst, interpretation shown as follows. The larger NMI or GS, the better the forecasts. We can find that ECMWF is the largest among the four products both in Figures 5a and 5b. In comparison, NCEP is the smallest in most of the six subfigures of Figure 5a and most of the six

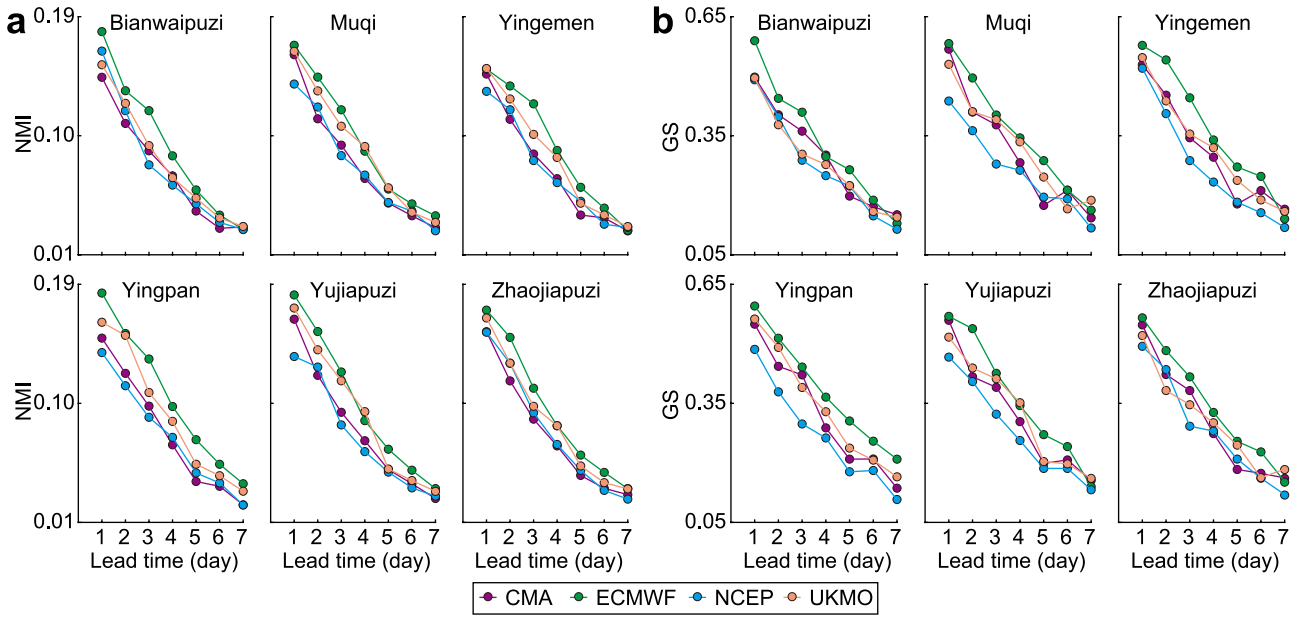

Figure 5. (a) Normalized mutual information (NMI) and (b) Gerrity score (GS) for the four forecast products China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP), and United Kingdom Meteorological Office (UKMO) at the six precipitation stations. The proposed score NMI and the traditional score GS were used to assess the comprehensive uncertainty of all categories.

subfigures of Figure 5b. (b) The performance of forecasts decreases with the increase of lead times and the downtrend is approximately linear. The conclusion (b) can help to estimate the performance of forecasts with a certain lead time if only limited data is available. Therefore, by comparing the results (a) and (b), we can get the conclusion that the NMI is as good as GS in terms of distinguishing the performance of different forecast products and capturing the changing patterns of uncertainties with lead times; and (c) Both NMI and GS are close to 0 when the lead time is +7 days. Therefore, the daily precipitation forecasts with a lead time more than 7 days almost have no practical meaning, which was also found by H. Wang et al. (2021) in the Huaihe River Basin, China.

4.3. Verification of the Uncertainty of a Certain Category

The proposed score NMI_k with the traditional scores IQR_k and Std_k were applied in the uncertainty assessment of forecast products for each category. The larger NMI_k or the smaller IQR_k and Std_k, the better performance of forecasts in the kth category. Figure 6 shows the NMI_k, IQR_k, and Std_k of four forecast products in category L1. Among the four products, the NMI_{L1} of ECMWF is the largest while the IQR_{L1} and Std_{L1} of ECMWF are the least for most of the lead times and precipitation stations. Therefore, for the three verification methods, ECMWF almost performs best at all the precipitation stations and lead times. Similarly, NCEP and UKMO perform worst as judged by NMI_{L1} and IQR_{L1}, respectively. However, no product performs worst according to Std_{L1} at all the precipitation stations and lead times. Therefore, the NMI_{L1} and IQR_{L1} show a slight advantage over Std_{L1} in distinguishing the forecasting performance among different forecast products.

In addition, we can find that the NMI_{L1} of the four products basically shows an obvious linear decreasing trend while both IQR_{L1} and Std_{L1} basically show an apparent increasing trend. However, there are some nearly horizontal segments in IQR_{L1}. The reason may be that IQR_{L1} cannot capture the change of the samples smaller than 25th percentile or larger than 75th percentile. The fluctuation of Std_{L1} may be caused by the large bias, which will be interpreted in Figure 9. Therefore, the NMI_{L1} shows a slight advantage over the IQR_{L1} and Std_{L1} on capturing the changing patterns of uncertainties with lead times.

The verification results show similarities for comprehensive uncertainty of all categories and uncertainty of category L1. For example, both the NMI and the NMI_{L1} show that ECMWF performs best and NCEP performs worst and the performance of forecasts decreases with the increase of lead times. It is explained in Table 4 that the

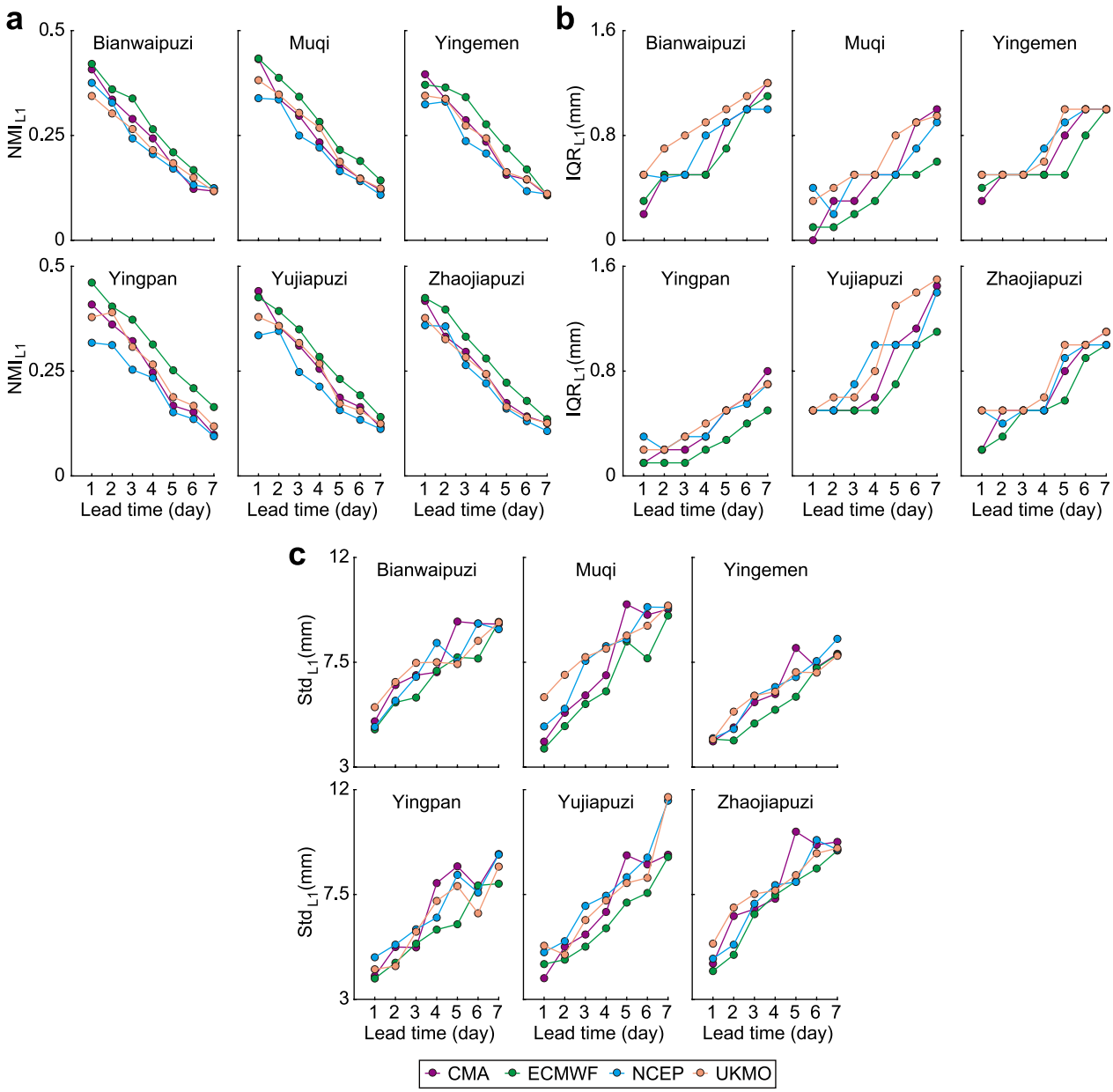

Figure 6. (a) NMI_{L1}, (b) IQR_{L1}, and (c) Std_{L1} for the four forecast products China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP), and United Kingdom Meteorological Office (UKMO) at the six precipitation stations. The proposed score NMI_{L1} with the traditional scores IQR_{L1} and Std_{L1} were applied in the uncertainty assessment of forecast products for category L1.

number of observations in category L1 (0–9.9 mm including no rain and light rain events) takes account of about 89% of all samples. Therefore, the similarities also verify the linkage between NMI and NMI_k in Equation 3.

Similarly, Figures 7 and 8 show the three verification methods in categories L2 and L3, respectively. The points E1 in Figure 7c and E2 in Figure 8b are the examples randomly selected from the outliers of Std_k and IQR_k, respectively. The NMI_{L2} and NMI_{L3} of four products basically show an increasing trend. The IQR_{L2} and IQR_{L3} basically show a decreasing trend and fluctuate larger than the NMI_{L2} and NMI_{L3}. However, the Std_{L2} and Std_{L3} fluctuate largest with lead times and show no obvious trend. Therefore, the NMI_k is better than the IQR_k and Std_k on capturing the changing patterns of uncertainties with lead times in categories L2 and L3.

In addition, we can find from NMI_{L2} and NMI_{L3} that CMA performs best in most of the seven lead times at the six precipitation stations. However, due to large fluctuations, it is hard to find the best products by using IQR_{L2},

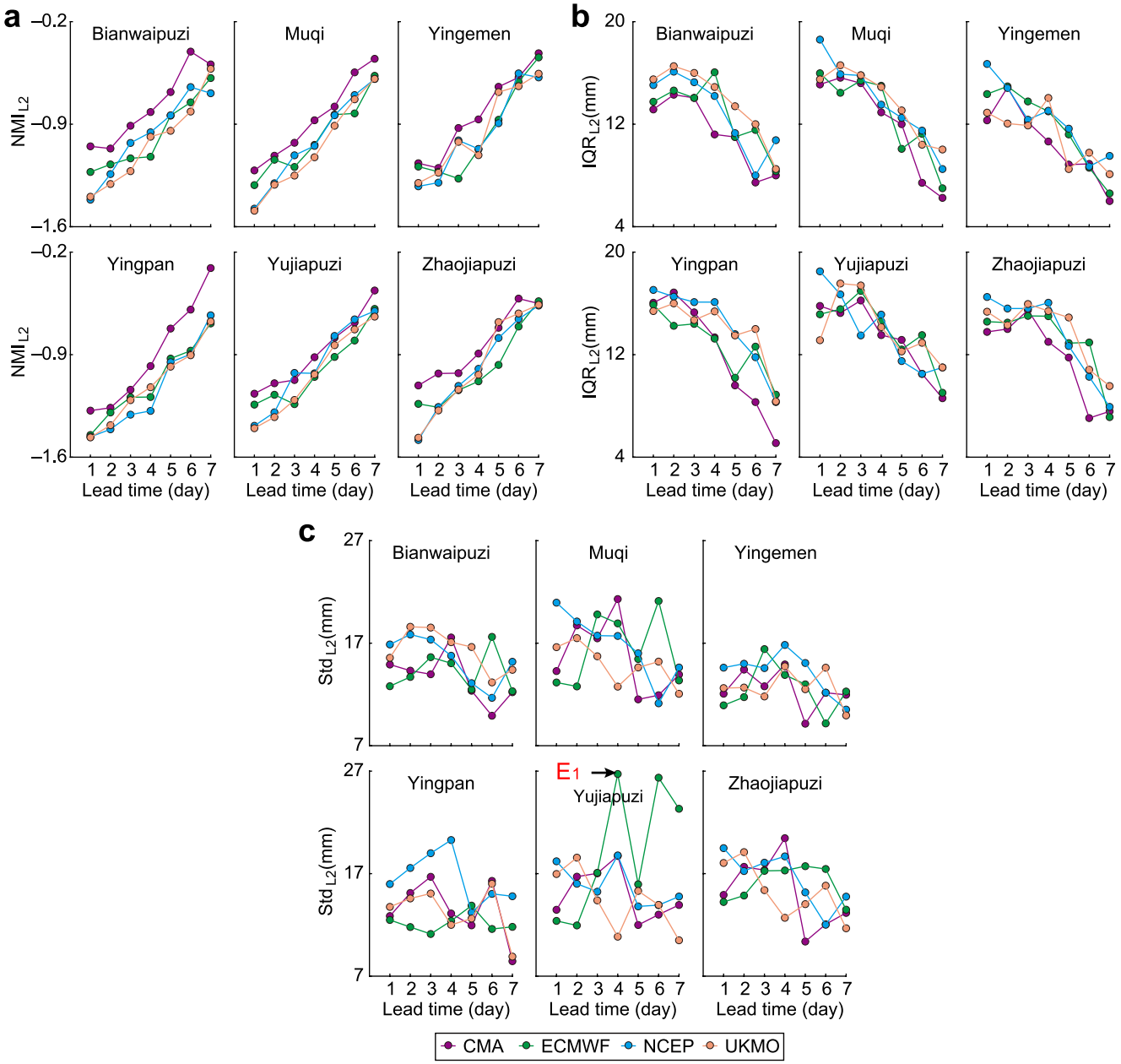

Figure 7. (a) NMI_{L2}, (b) IQR_{L2}, and (c) Std_{L2} for the four forecast products China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP), and United Kingdom Meteorological Office (UKMO) at the six precipitation stations. The proposed score NMI_{L2} with the traditional scores IQR_{L2} and Std_{L2} were applied in the uncertainty assessment of forecast products for category L2. E_1 is an example of outlier in panel (c).

IQR_{L3}, Std_{L2}, and Std_{L3} (especially for Std_{L2} and Std_{L3}). Therefore, the NMI_k is better than the IQR_k and Std_k in distinguishing the forecasting performance among different forecast products in categories L2 and L3.

From Figures 7 and 8, we can find that the outliers contribute significantly to the fluctuation of Std_k and IQR_k. The outlier E_1 in Figure 7c is used to explain the fluctuation of Std_k. The value of the point E_1 is abnormal and obviously larger than the values of other points in Figure 7c. The explanation can be found in Figure 9a, which shows the observed precipitation when the corresponding forecasts from ECMWF with a lead time of +4 days at Yujiapuzi belong to category L2. In Figure 9a, the point F is an outlier with an extreme value of 311.5 mm. If the point F is removed, the Std_{L2} will decrease significantly from 26.8 to 15.7 mm. Thus, being very sensitive to extreme bias is one of the critical reasons for the significant fluctuation of Std_k. In addition, the fluctuation of

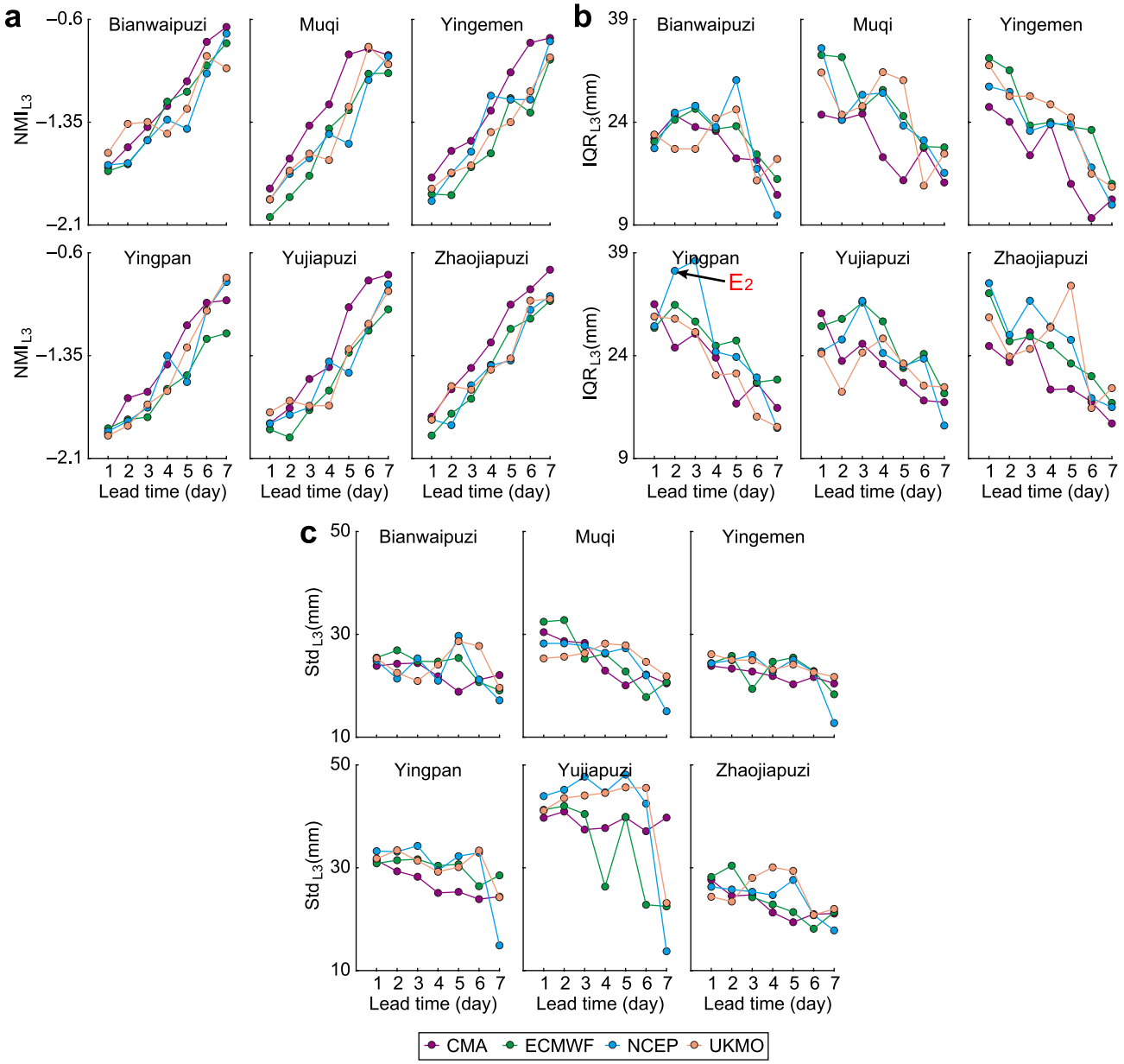

Figure 8. (a) NMI_{L3}, (b) IQR_{L3}, and (c) Std_{L3} for the four forecast products China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP), and United Kingdom Meteorological Office (UKMO) at the six precipitation stations. The proposed score NMI_{L3} with the traditional scores IQR_{L3} and Std_{L3} were applied in the uncertainty assessment of forecast products for category L3. E_2 is an example of outlier in panel (b).

Std_{L1} is smaller than that of Std_{L2} and Std_{L3}. One reason is that the extreme values of the observations are easier to occur when the corresponding forecasts belong to the high categories (L2 and L3) than those in the low category (L1). Another reason can be found by Equation 14. The larger the sample size m, the more stable the Std_k. In Table 4, the sample size of observed precipitation in category L1 is much larger than that in categories L2 and L3. Thus, the fluctuation of Std_{L1} is smaller than that of Std_{L2} and Std_{L3}.

Similarly, we can use the outlier E_2 in Figure 8b to explain the fluctuation of IQR_k. The value of the point E_2 is abnormal and obviously larger than most values of the other points in Figure 8b. The explanation can be found by Figure 9b, which shows the cumulative frequency of the observed precipitation when the corresponding forecasts from NCEP with a lead time of +2 days at Yingpan belong to category L3. In Figure 9b, points C and D are the 75th and 25th percentiles of the values of the observed precipitation, respectively. Using the linear interpretation

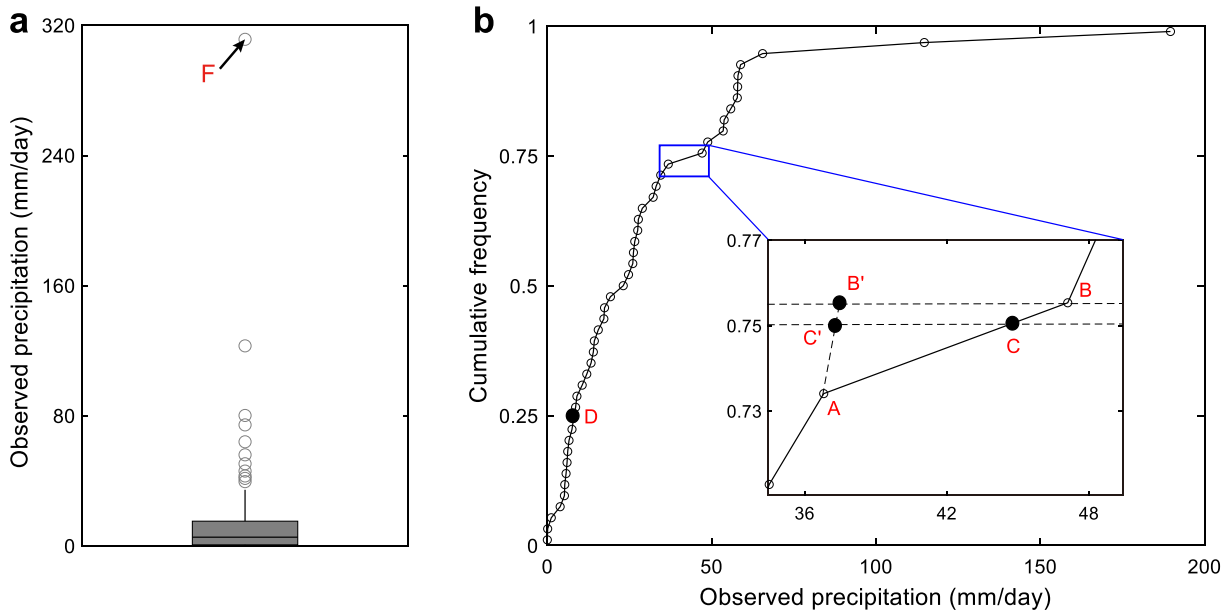

Figure 9. Interpretation for the outliers of (a) E1 (in Figure 7) and (b) E2 (in Figure 8), respectively. (a) The observed precipitation when the corresponding forecasts from European Centre for Medium-Range Weather Forecasts (ECMWF) with a lead time of +4 days at Yujiapuzi belong to category L2; and (b) the cumulative frequency of the observed precipitation when the corresponding forecasts from National Centers for Environmental Prediction (NCEP) with a lead time of +2 days at Yingpan belong to category L3. A and B are the nearest left and right points of C, respectively. C and D are the 75th and 25th percentiles of the values of the observed precipitation, respectively. B moves horizontally a random distance to B′. C′ is the intersection of line AB′ and horizontal line CC′. F is an outlier.

method, the coordinate of C can be calculated by the coordinates of A and B, which are the nearest left and right points of C, respectively. To test the stability of IQR_k, we can change the value of the observation at B by moving B to a random position B′. Then, the position of C will change to C′. The coordinates of these points are: A (36.8, 0.745), B (47.1,0.766), B′ (37.5,0.766), C (44.5, 0.750), C′ (37.3, 0.750), and D (8.2, 0.250). It means that if only one of these values of observed precipitation changes from 47.1 to 37.5 mm, IQR_{L3} will change significantly from 36.3 mm (44.5 minus 8.2) to 29.1 mm (37.3 minus 8.2). From Figure 9b, we can also find that the IQR_{L3} would be less stable if the points around C are sparser. Therefore, the smaller the sample size, the less accurate the IQR_k and the higher fluctuation of IQR_k with lead times. From Table 4, we find that the sample size of observed precipitation in categories L1, L2, and L3 are around 2,000, 160, and 100, respectively. Therefore, among these three categories, IQR_k fluctuate largest in category L3 and least in category L1.

4.4. Experiments for Evaluating the Resistivity of Verification Methods to Extreme Bias

In Section 4.3, through the example, we only showed the resistivity of Std_k to extreme biases. Here, we set up more experiments to evaluate the resistivity of all verification methods, namely GS, NMI, NMI_k, IQR_k, and Std_k. The observed precipitation data used for the resistivity experiments are obtained from the Muqi station. The corresponding forecasts are obtained from ECMWF with a lead time +1 day. We randomly selected 1 day for these experiments, such as 13 August 2017. The values of observation and forecast on that day are 26.5 and 2.3 mm, respectively. The value of the forecast (2.3 mm) belongs to category L1. Therefore, we took category L1 as an example to compare the resistivity of NMI_k, IQR_k, and Std_k. The variation of the five verification methods can be analyzed by changing the bias from −24.2 mm (2.3 minus 26.5) to different extreme values in different experiments. To make it easier to compare the verification results between different experiments and the original, the value of the forecast on that day in different experiments is set to the same value as the original, that is, 2.3 mm. In these experiments, thus, the value of the hypothetic observation on that day varies with the bias. The parameters and corresponding results of these experiments are shown in Table 8. GS-p, NMI-p, NMI_{L1}-p, IQR_{L1}-p and Std_{L1}-p represent the corresponding percentage changes of GS, NMI, NMI_{L1}, IQR_{L1}, and Std_{L1}, respectively.

Table 8
Parameters and Results of Resistivity Experiments of Verification Methods to Extreme Bias

Index	Original	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Bias (mm)	−24.2	−50.0	−100.0	−150.0	−200.0
Forecast (mm)	2.3	2.3	2.3	2.3	2.3
Observation (mm)	26.5	52.3	102.3	152.3	202.3
GS	0.582	0.582	0.582	0.582	0.582
GS-p (%)	–	0	0	0	0
NMI	0.168	0.169	0.172	0.173	0.178
NMI-p (%)	–	0	2	3	6
NMI_{L1}	0.433	0.433	0.434	0.434	0.445
NMI_{L1}-p (%)	–	0	0	0	3
IQR_{L1} (mm)	0.1	0.1	0.1	0.1	0.1
IQR_{L1}-p (%)	–	0	0	0	0
Std_{L1} (mm)	3.8	3.9	4.4	5.1	5.9
Std_{L1}-p (%)	–	4	16	35	57

Note. GS-p, NMI-p, NMI_{L1}-p, IQR_{L1}-p, and Std_{L1}-p represent the corresponding percentage changes of GS, NMI, NMI_{L1}, IQR_{L1}, and Std_{L1}, respectively.

4.4.1. Resistivity Analyses of GS and NMI

It is noted that the GS keeps unchanged. The reason is that the normalized contingency table (as shown in Data Availability Statement) to calculate the GS keeps unchanged in the four experiments. Therefore, the GS nearly brings no information about the change of the extreme bias. The NMI increases slightly with the increase of the absolute value of the extreme bias. According to the binning method M1 in Table 6, the bin width is calculated by the σ of the sample, which is affected by the bias. Therefore, the change of bias indirectly leads to the change of NMI.

4.4.2. Resistivity Analyses of NMI_k, IQR_k, and Std_k

The NMI_{L1} changes no more than 0.001 in the first three experiments and only increases 3% in Experiment 4. The reason for the change of NMI_{L1} is the same as NMI. However, the NMI_{L1} is less sensitive than NMI to extreme bias and nearly keeps unchanged in the first three experiments. IQR_{L1} stays unchanged in the four experiments. The reason is shown as follows. IQR_{L1} is only affected by the 75th and 25th percentiles of observed precipitation when corresponding forecasts belong to category L1. However, the 75th and 25th percentiles of observed precipitation keep unchanged because the extreme values of the hypothetic observed precipitation in the four experiments are larger than the 75th percentile of observed precipitation when forecasts belong to category L1. Similar to GS, the IQR_{L1} nearly brings no information about the extreme bias. In comparison, the Std_{L1} changes much with the increase of the extreme biases, which is up to 57% in Experiment 4. Thus, the Std_{L1} is the most sensitive to extreme biases among the three verification methods.

The access to the details, including the data and MATLAB code, for the experiments are shown in Data Availability Statement.

5. Discussion

5.1. The Comparison of Verification Results at Different Precipitation Stations

The comprehensive uncertainty and uncertainty of a certain category of precipitation forecasts in DRDB were assessed in the study. We compared the performance of forecasts at different precipitation stations by the five scores, namely NMI, GS, NMI_k, IQR_k, and Std_k. As indicated in the results, the climatologies for the six precipitation stations in the DRDB differ slightly given their spatial scale, which is consistent with Figure 2a. In particular, the five scores at Muqi and Zhaojiapuzi, two precipitation stations with close locations (shown in Figure 1a), have presented similar results.

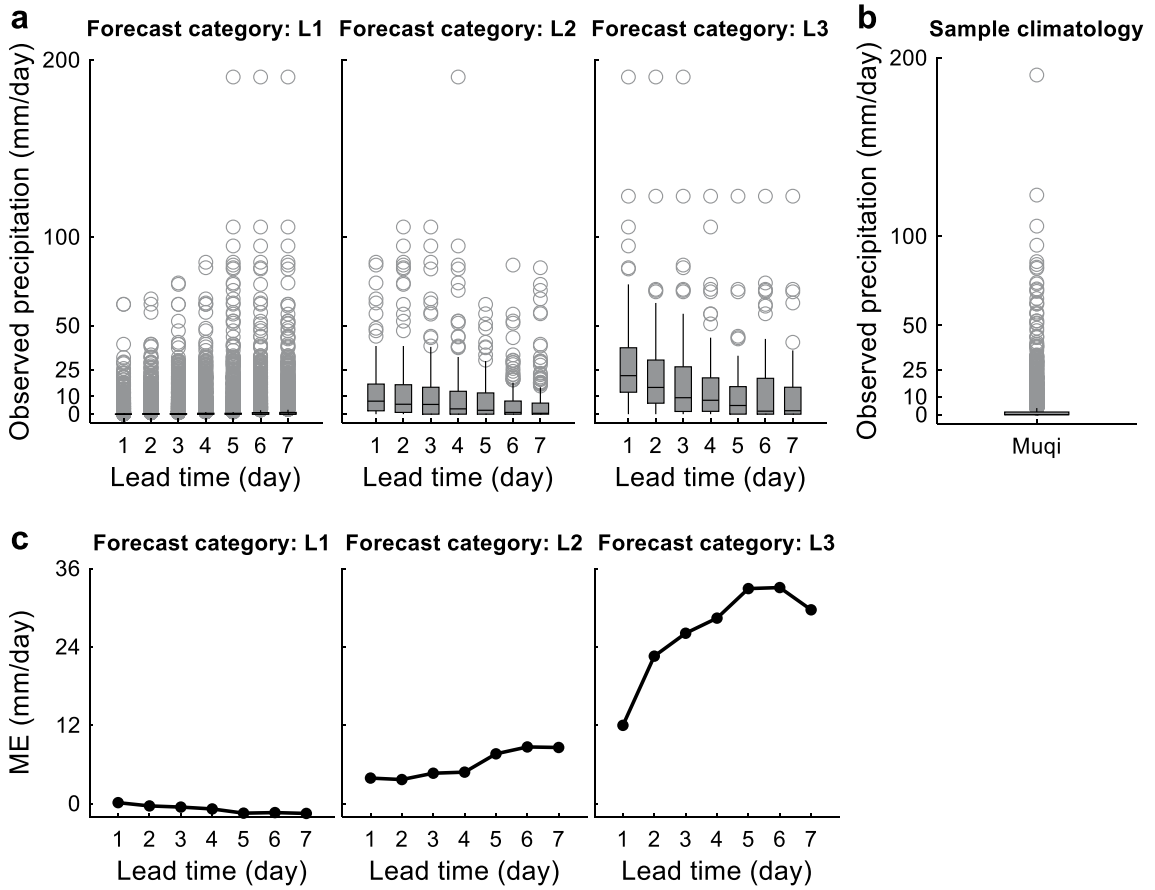

Figure 10. Interpretation for the changing patterns of uncertainties with lead times in each category. (a) The box and whisker plots for observed precipitation when the corresponding forecasts from China Meteorological Administration (CMA) with seven lead times at Muqi belong to one certain category; (b) the box and whisker plot for observed precipitation at Muqi (sample climatology); and (c) the mean error of forecasts corresponding to panel (a).

5.2. Interpretation for the Changing Patterns of Uncertainties With Lead Times in Each Category

According to NMI_k, IQR_k, and Std_k for categories L2 and L3, the uncertainty of forecasts decreases with the increase of lead times, which is different from the results of NMI, GS and the three verification methods for category L1. However, as mentioned in the introduction, the uncertainty assessed in the study is only one of the attributes of the performance of forecasts. Therefore, we cannot draw the conclusion from Figures 7 and 8 that the overall performance of forecasts in categories L2 and L3 decreases with the increase in lead times, which can be explained by Figure 10. The three subfigures in Figure 10a (denoted as $SubF^a_{L1}$, $SubF^a_{L2}$, and $SubF^a_{L3}$, respectively) are the box and whisker plots for observed precipitation when the corresponding forecasts from CMA with seven lead times at Muqi belong to category L1, L2, and L3, respectively. It should be noted that the values of the points in $SubF^a_{L1}$, that is, the values of observed precipitation, may not belong to category L1 because forecast errors are inevitable. For example, the values of observed precipitation belong to category L3 when the corresponding points are above 25 mm. Likewise, the values of the points in $SubF^a_{L2}$ and $SubF^a_{L3}$ do not necessarily belong to categories L2 and L3, respectively. Figure 10b is the box and whisker plot for all observed precipitation at Muqi, which shows the empirical distribution function of observable precipitation based on a sample of past observations at Muqi, that is, sample climatology. The three subfigures in Figure 10c (denoted as $SubF^c_{L1}$, $SubF^c_{L2}$, and $SubF^c_{L3}$, respectively) show the ME of forecasts corresponding to Figure 10a. In Figure 10a, most of the samples are smaller than 10 mm and concentrated near 0 mm. In addition, the ME in $SubF^c_{L1}$ are close to 0 mm. Therefore, no obvious bias exists in category L1. However, most of the samples are smaller than 10 mm in $SubF^a_{L2}$ and smaller than 25 mm in $SubF^a_{L3}$. In addition, the ME in $SubF^c_{L2}$ and $SubF^c_{L3}$ are obviously larger than 0 mm. Therefore, most of the forecasts in categories L2 and L3 are larger than the observed precipitation. What is more, the biases increase with the increase of lead times. With the increase of lead times, the samples in $SubF^a_{L1}$ tend to be

Table 9
Bin Width to Calculate the Entropy of Each Precipitation Station by Five Fixed Width Binning Methods M1 to M5 (mm)

	Precipitation station					
Binning method ID	Bianwaipuzi	Muqi	Yingemen	Yingpan	Yujiapuzi	Zhaojiapuzi
M1	2.7	2.8	2.4	2.7	3.2	2.8
M2	10.2	14.6	9.5	14.6	24.0	11.2
M3	11.0	15.9	10.3	15.8	26.0	12.2
M4	2.8	4.1	2.6	4.0	6.6	3.1
M5	0.1	0.1	0.1	0.1	0.1	0.1

Note. The details of the binning methods M1 to M5 is shown in Table 6.

sparser while the samples in $SubF^{a}_{L2}$ and $SubF^{a}_{L3}$ tend to be more concentrated. The more concentrated the sample, the smaller the uncertainty. Therefore, the uncertainty of forecasts increases with the increase of lead times in forecast category L1 and decreases with the increase of lead times in forecast categories L2 and L3, which is consistent with the results of NMI_k, IQR_k, and Std_k in these three categories (Figures 6–8). Compared with the sample climatology in Figure 10b, the samples in $SubF^{a}_{L1}$ are more concentrated while the samples in $SubF^{a}_{L2}$ and $SubF^{a}_{L3}$ are less concentrated. According to entropy theory (Singh, 2013, p. 126), the concentration of samples in Figures 10a and 10b can be assessed by $H(O|F_k)$ and $H(O)$ in Equation 4, respectively. Thus, $H(O|F_{L1})$ is smaller than $H(O)$ while $H(O|F_{L2})$ and $H(O|F_{L3})$ are larger than $H(O)$. Therefore, by Equation 4, NMI_{L1} is larger than 0 while NMI_{L2} and NMI_{L3} are smaller than 0, which is consistent with the results of Figures 6a, 7a and 8a, respectively. With the increase of lead times, the distribution of the samples in Figure 10a tends to be that in Figure 10b. Thus, NMI_k in all categories tend to be 0 and the forecasts tend to be independent from observed precipitation which is also in line with common sense.

Based on the analysis shown above, we can also find that bias analysis is necessary before verifying the uncertainties in precipitation forecasts.

5.3. Binning Method Comparison and Selection for the Proposed Approach

As shown in Section 3, we introduced five fixed width binning methods and one non-fixed binning method. Binning method selection is important because an inappropriate method would bring extra errors. Therefore, it is necessary to carefully compare the bin width and the verification results by different binning methods.

5.3.1. Comparison of Bin Width by Using Different Binning Methods

With Table 6, Equation 5 and the data of O at each station, the bin width by five fixed width binning methods can be obtained, as shown in Table 9.

From Table 9, we can find that the bin widths from different binning methods follow this order: M5 < M1 < M4 < M2 < M3. The bin width from M5 at different precipitation stations are all equal to 0.1 mm, which is the resolution of observed precipitation. The bin widths from M1 are similar to that of M4 and the bin widths from M2 are similar to that of M3. The difference of bin widths at different precipitation stations from M1 are smaller than that from M2, M3, and M4. As shown in Table 6, the bin width by M1, M2, M3, and M4 can be calculated by certain equations while M5 is based on maximum a posteriori estimation and Bayes' theorem. Therefore, it is hard to analyze the relationship between M5 and other four binning methods. In this section, we analyzed the factors affecting the order of the bin width calculated by M1, M2, M3, and M4 in Table 6, shown as follows.

According to Table 6 and Equation 5, we can obtain

$$W_{M1} = 3.49 \times \sigma \times S^{-\frac{1}{3}} \tag{15}$$

$$W_{M2} = \frac{R}{\mathrm{Int}\left(1 + \log_2 S\right)} \tag{16}$$

Table 10
The Ratio of the Bin Width by Binning Methods M1 and M4 at Each Precipitation Station

Precipitation station	Bianwaipuzi	Muqi	Yingemen	Yingpan	Yujiapuzi	Zhaojiapuzi
σ (mm)	10.1	10.7	9.0	10.2	12.1	10.6
R (mm)	132.0	190.4	123.5	189.6	311.5	145.8
$\frac{W_{M1}}{W_{M4}}$ (by Equation 19)	0.96	0.70	0.92	0.68	0.49	0.91
$\frac{W_{M1}}{W_{M4}}$ (by Table 9)	0.96	0.68	0.92	0.68	0.48	0.90

Note. σ: the standard deviation of the distribution of the observed precipitation; R: the range of distribution of the observed precipitation; W_{M1}: the bin width of the binning methods M1; W_{M4}: the bin width of the binning methods M4.

$$W_{M3} = \frac{R}{\mathrm{Int}\left(1 + 1.33 \times \log_e S\right)} \tag{17}$$

$$W_{M4} = \frac{R}{\mathrm{Int}(\sqrt{S})} \tag{18}$$

where W_{M1}, W_{M2}, W_{M3}, and W_{M4} are the bin width of the binning methods M1, M2, M3, and M4, respectively.

5.3.1.1. The Relationship of the Bin Width Between M1 and M4

With Equations 15 and 18, the ratio between W_{M1} and W_{M4} are shown as follows

$$\frac{W_{M1}}{W_{M4}} = \frac{3.49 \times \sigma \times S^{-\frac{1}{3}}}{\frac{R}{\mathrm{Int}(\sqrt{S})}} \approx \frac{3.49 \times \sigma \times S^{\frac{1}{6}}}{R} \tag{19}$$

In Equation 19, we can find that $\frac{W_{M1}}{W_{M4}}$ is affected by S, σ, and R. In the study, S equals to 2,208, that is the same for the six rainfall stations, as shown in Table 4. σ and R differ in different precipitation stations. Therefore, it is necessary to calculate σ and R to obtain $\frac{W_{M1}}{W_{M4}}$, as shown in Table 10. In Table 10, we can find that $\frac{W_{M1}}{W_{M4}}$ calculated by Equation 19 (approximate values) is similar to that by Table 9 (accurate values). In Table 10, We can also find that $\frac{W_{M1}}{W_{M4}} < 1$ at the six precipitation stations. Thus, we can obtain

$$W_{M1} < W_{M4} \tag{20}$$

which is tenable at the six precipitation stations.

5.3.1.2. The Relationship of the Bin Width Between M2 and M3

With Equations 16 and 17, the ratio between W_{M2} and W_{M3} are shown as follows

$$\frac{W_{M2}}{W_{M3}} = \frac{\frac{R}{\mathrm{Int}\left(1+\log_2 S\right)}}{\frac{R}{\mathrm{Int}\left(1+1.33\times\log_e S\right)}} \approx \frac{1 + 0.92 \times \log_2 S}{1 + \log_2 S} \tag{21}$$

At the six precipitation stations, $\frac{W_{M2}}{W_{M3}} \approx \frac{1+0.92\times\log_2 2{,}208}{1+\log_2 2{,}208} = 0.93$ (the accurate value of $\frac{W_{M2}}{W_{M3}}$ can also be calculated by Table 9). Therefore, W_{M2} is a little smaller than W_{M3} at the six precipitation stations, as shown in Table 9.

In Equation 21, we can find that $\frac{W_{M2}}{W_{M3}}$ is only affected by S. Therefore, with Equation 21, we can get two Equations 22 and 23, which are also tenable in other basins.

$$0.92 < \frac{W_{M2}}{W_{M3}} < 1 \tag{22}$$

$$W_{M2} < W_{M3} \tag{23}$$

5.3.1.3. The Relationship of the Bin Width Between M4 and M2

With Equations 18 and 16, the ratio between W_{M4} and W_{M2} are shown as follows

$$\frac{W_{M4}}{W_{M2}} = \frac{\frac{R}{\mathrm{Int}(\sqrt{S})}}{\frac{R}{\mathrm{Int}(1+\log_2 S)}} \approx \frac{1+\log_2 S}{\sqrt{S}} \tag{24}$$

At the six precipitation stations, $\frac{W_{M4}}{W_{M2}} \approx \frac{1+\log_2 2{,}208}{\sqrt{2{,}208}} = 0.26$ (the accurate value of $\frac{W_{M4}}{W_{M2}}$ can also be calculated by Table 9). Therefore, W_{M4} is much smaller than W_{M2} at the six precipitations, as shown in Table 9.

In Equation 24, we can find that $\frac{W_{M4}}{W_{M2}}$ is also only affected by S, which should be large enough to be representative. Therefore, we can get Equations 25 and 26, which are also tenable in other basins.

$$\lim_{S\to+\infty} \frac{W_{M4}}{W_{M2}} = 0 \tag{25}$$

$$W_{M4} < W_{M2} \tag{26}$$

Finally, according to Equations 20, 23 and 26, we can get

$$W_{M1} < W_{M4} < W_{M2} < W_{M3} \tag{27}$$

which are tenable at the six precipitation stations.

5.3.2. Binning Method Selection for the Proposed Approach

The bin width may show much difference if taking different binning methods and it is necessary to find which one is suitable for the study. In the following, we will use the first two criteria (shown at the beginning of Section 4) to select a proper binning method.

With the bin width, NMI and NMI_k by different binning methods at Muqi can be calculated as one example and are shown in Figure 11.

The range of NMI or NMI_k with different lead times follows this order: (M1 ≈ M4 ≈ M5) < (M0 ≈ M2 ≈ M3) except that the range of NMI_{L3} by M0 is similar to that of M1, M4, and M5. The width of three bins by M0 are 10 mm (10 minus 0), 15 mm (25 minus 10) and + ∞ (≥25). With the order of the bin width by using different binning methods, we can find that the range of NMI or NMI_k with different lead times tends to increase with the increase of bin width. The larger the range of NMI or NMI_k, the easier to distinguish the forecast performance with different lead times. However, a large range of bin width also brings significant fluctuation to NMI and NMI_k which makes it hard to capture the changing patterns of uncertainties with lead times and distinguish the forecasting performance among different forecast products.

In Figure 11c, it is difficult to distinguish the performance of different products by applying the binning methods M2 and M3. In Figure 11d, it is hard to capture the changing patterns of uncertainties with lead times by using the binning method M0. Therefore, binning methods M1, M4, and M5 show similar performance and are better than M0, M2, and M3. In addition, among these five binning methods, M1, which was proposed by Scott (1979), is most frequently used (Loritz et al., 2019). Because M1 performs well, it is applied in the study. It should be noted that the most proper binning method may vary in other basins.

Extreme precipitation forecasts are vital for water resource management (e.g., flood control). Due to the small sample size, however, the forecasting precipitation greater than the medium precipitation as defined by the CMA are merged into category L3 in this study. Therefore, the uncertainty verification for the extreme precipitation forecasts is conjectural as represented in this study. For future work, it is expected that the proposed method would be implemented in different climate zones with more case studies (e.g., large basins) for comparing the performance of precipitation forecasting. In addition, more research work should be focused on the applicability

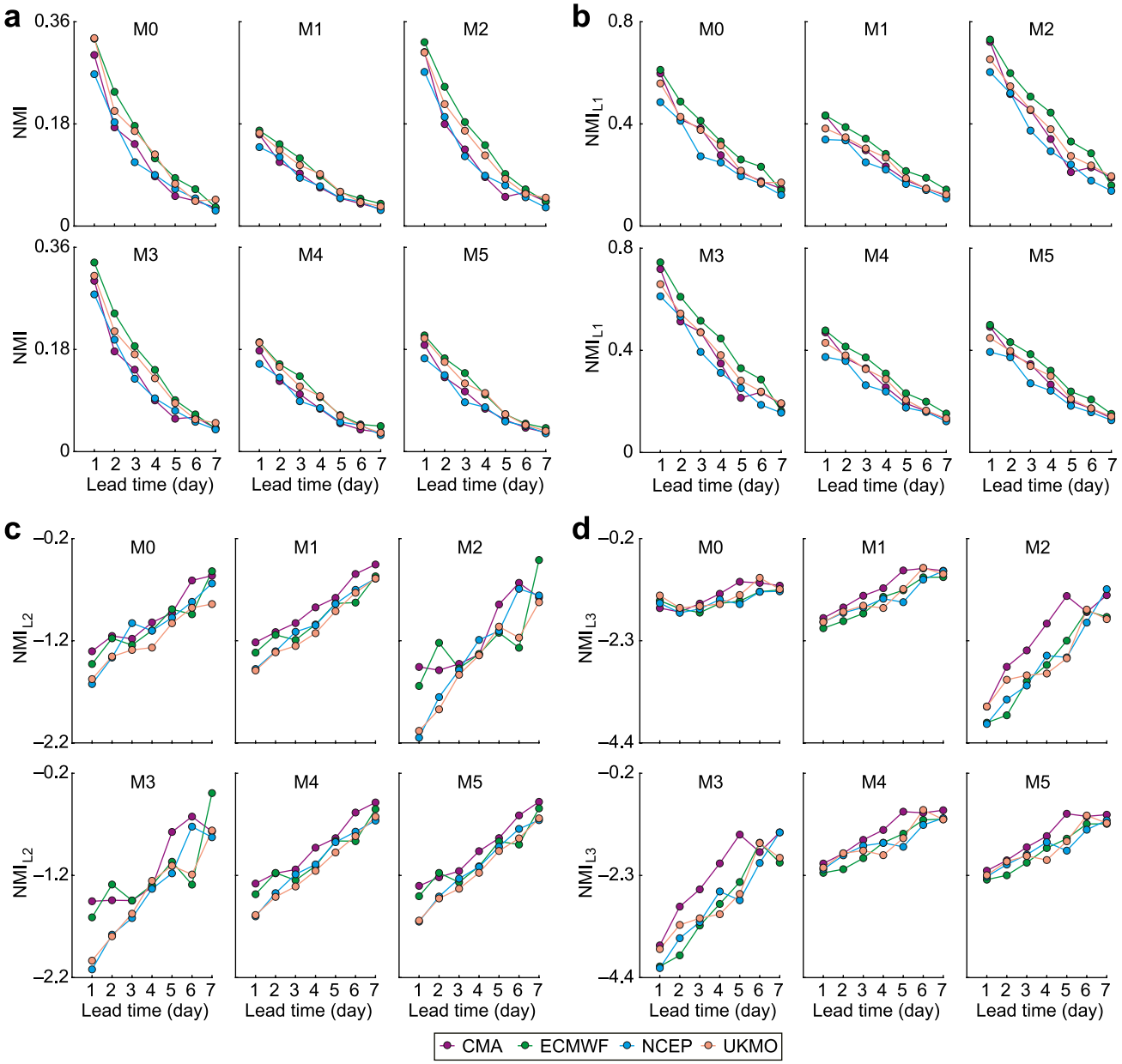

Figure 11. (a) Normalized mutual information (NMI), (b) NMI_{L1}, (c) NMI_{L2}, and (d) NMI_{L3} for the four forecast products China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP), and United Kingdom Meteorological Office (UKMO) by six binning methods at Muqi.

of the proposed scores in other hydrometeorological forecasts, such as deterministic multi-category forecasts of temperature and streamflow.

6. Conclusions

Verification for the deterministic multi-category precipitation forecasts is important for end-users. This study proposed a new approach using two mutual information theory-based scores, namely NMI and NMI_k (the decomposition of NMI in the kth category), for assessing the comprehensive uncertainty and the uncertainty for a certain category, respectively in deterministic multi-category precipitation forecasting. The purpose of these two scores is for the verification of forecasting products. Specifically, the comprehensive uncertainty is defined as the average reduction in uncertainty about the observations resulting from the use of a predictive model to

provide all categories forecasts; the uncertainty of a certain category is defined as the reduction in uncertainty about the observations resulting from the use of a predictive model to provide a certain category forecast. These two proposed scores can make full use of the forecasting and observed data information, which are resistant to extreme biases and suitable for comparing the performance of forecasts under different climatologies. In addition, the proposed approach establishes the linkage between the comprehensive performance of all categories of forecast (NMI) and the forecast performance for a certain category (NMI_k).

By applying the mutual information theory-based approach and traditional verification methods in the DRDB, we conclude the following:

1. In terms of capturing the changing patterns of uncertainties with lead times and distinguishing the forecasting performance among different forecast products, the NMI_k shows a better performance than the other two reference verification methods IQR_k (the IQR in the kth category) and Std_k (the Std in the kth category), while NMI shows a similar performance with the reference verification method GS.
2. The NMI and NMI_k are resistant to extreme biases.
3. The difference between the NMI and NMI_k using different binning methods indicates that a careful choice of bin width is needed.
4. The large anomalies of forecasting performance in categories L2 and L3 show that a bias analysis is necessary before verifying the uncertainties in precipitation forecasts.

Acronyms

BR	bias ratio
CMA	China Meteorological Administration
DRDB	Dahuofang Reservoir Drainage Basin
DS	divergence score
ECMWF	European Centre for Medium-Range Weather Forecasts
GS	Gerrity score
IQR	interquartile range
ME	mean error
NCEP	National Centers for Environmental Prediction
NMI	normalized mutual information
NWP	numerical weather prediction
ORGG	octahedral reduced Gaussian grid
PC	proportion correct
POD	probability of detection
Std	standard deviation
TIGGE	Interactive Grand Global Ensemble
UKMO	United Kingdom Meteorological Office

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Data and code used in this study are available at https://doi.org/10.4211/hs.48c6a00bb6c449afbe33b67250cd1ae7.

Acknowledgments
We really appreciate the time and efforts that the editors and reviewers have dedicated on the manuscript, which help to improve the quality of this manuscript significantly. In addition, we would also like to thank Dr. Lei Zhang, who helps us polish the figures in the manuscript. The work was supported by the National Natural Science Foundation of China (Grants 51779030 and 52079015), the BMSTC | Beijing Science and Technology Planning Project (Grant Z191100006919002), and the University of Glasgow CoSS Strategic Research Fund (Grant PO20028963). The first author gratefully acknowledges the financial support provided by the China Scholarship Council (Grant 201906060080).

References

Ahrens, B., & Walser, A. (2008). Information-based skill scores for probabilistic forecasts. *Monthly Weather Review*, *136*(1), 352–363. https://doi.org/10.1175/2007MWR1931.1

Alfonso, L., Lobbrecht, A., & Price, R. (2010). Information theory-based approach for location of monitoring water level gauges in polders. *Water Resources Research*, *46*(3), W03528. https://doi.org/10.1029/2009WR008101

Babel, M. S., Badgujar, G. B., & Shinde, V. R. (2015). Using the mutual information technique to select explanatory variables in artificial neural networks for rainfall forecasting: MI technique in rainfall forecasting. *Meteorological Applications*, *22*(3), 610–616. https://doi.org/10.1002/met.1495

Bradley, A. A., Demargne, J., & Franz, K. J. (2016). Attributes of forecast quality. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), *Handbook of hydrometeorological ensemble forecasting* (pp. 1–44). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40457-3_2-1

Brier, G. W., & Allen, R. A. (1951). Verification of weather forecasts. In H. R. Byers, H. E. Landsberg, H. Wexler, B. Haurwitz, A. F. Spilhaus, H. C. Willett, et al. (Eds.), *Compendium of meteorology* (pp. 841–848). American Meteorological Society. https://doi.org/10.1007/978-1-940033-70-9_68

Brooks, H. E., & Doswell, C. A. (1996). A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Weather and Forecasting*, *11*(3), 288–303. https://doi.org/10.1175/1520-0434(1996)011<0288:ACOMOA>2.0.CO;2

Brown, B. G., & Murphy, A. H. (1987). Quantification of uncertainty in fire-weather forecasts: Some results of operational and experimental forecasting programs. *Weather and Forecasting*, *2*(3), 190–205. https://doi.org/10.1175/1520-0434(1987)002<0190:QOUIFW>2.0.CO;2

Cai, C., Wang, J., & Li, Z. (2019). Assessment and modelling of uncertainty in precipitation forecasts from TIGGE using fuzzy probability and Bayesian theory. *Journal of Hydrology*, *577*, 123995. https://doi.org/10.1016/j.jhydrol.2019.123995

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., et al. (2008). Forecast verification: Current status and future directions. *Meteorological Applications*, *15*(1), 3–18. https://doi.org/10.1002/met.52

Dance, S., Ballard, S., Bannister, R., Clark, P., Cloke, H., Darlington, T., et al. (2019). Improvements in forecasting intense rainfall: Results from the FRANC (Forecasting Rainfall Exploiting New Data Assimilation Techniques and Novel Observations of Convection) project. *Atmosphere*, *10*(3), 125. https://doi.org/10.3390/atmos10030125

Davis, M., Lowe, R., Steffen, S., Doblas-Reyes, F., & Rodó, X. (2016). Barriers to using climate information: Challenges in communicating probabilistic forecasts to decision-makers. In J. L. Drake, Y. Y. Kontar, J. C. Eichelberger, T. S. Rupp, & K. M. Taylor (Eds.), *Communicating climate-change and natural hazard risk and cultivating resilience* (Vol. 45, pp. 95–113). Springer International Publishing. https://doi.org/10.1007/978-3-319-20161-0_7

DelSole, T., & Tippett, M. K. (2007). Predictability: Recent insights from information theory: Predictability. *Reviews of Geophysics*, *45*(4). https://doi.org/10.1029/2006RG000202

Dorninger, M., Ghelli, A., & Lerch, S. (2020). Editorial: Recent developments and application examples on forecast verification. *Meteorological Applications*, *27*(4), e1934. https://doi.org/10.1002/met.1934

Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Nurmi, P., Gill, P., et al. (2013). Progress and challenges in forecast verification: Progress challenges in forecast verification. *Meteorological Applications*, *20*(2), 130–139. https://doi.org/10.1002/met.1392

Economou, T., Stephenson, D. B., Rougier, J. C., Neal, R. A., & Mylne, K. R. (2016). On the use of Bayesian decision theory for issuing natural hazard warnings. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *472*(2194), 20160295. https://doi.org/10.1098/rspa.2016.0295

Fundel, V. J., Fleischhut, N., Herzog, S. M., Göber, M., & Hagedorn, R. (2019). Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users. *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 210–231. https://doi.org/10.1002/qj.3482

Gencaga, D., Knuth, K., & Rossow, W. (2015). A recipe for the estimation of information flow in a dynamical system. *Entropy*, *17*(1), 438–470. https://doi.org/10.3390/e17010438

Gilleland, E., Pappenberger, F., Brown, B., Ebert, E., & Richardson, D. (2016). Verification of meteorological forecasts for hydrological applications. In Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), *Handbook of hydrometeorological ensemble forecasting* (pp. 1–30). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40457-3_4-1

Gong, W., Gupta, H. V., Yang, D., Sricharan, K., & Hero, A. O. (2013). Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach: Estimating epistemic and aleatory uncertainties. *Water Resources Research*, *49*(4), 2253–2273. https://doi.org/10.1002/wrcr.20161

Gong, W., Yang, D., Gupta, H. V., & Nearing, G. (2014). Estimating information entropy for hydrological data: One-dimensional case. *Water Resources Research*, *50*(6), 5003–5018. https://doi.org/10.1002/2014WR015874

Hao, Z., & Singh, V. P. (2013). Entropy-based method for extreme rainfall analysis in Texas: Extreme rainfall analysis IN Texas. *Journal of Geophysical Research: Atmospheres*, *118*(2), 263–273. https://doi.org/10.1029/2011JD017394

Hughes, G., & McRoberts, N. (2014). The structure of diagnostic information. *Australasian Plant Pathology*, *43*(3), 267–286. https://doi.org/10.1007/s13313-013-0267-2

Hughes, G., McRoberts, N., & Burnett, F. J. (2017). Resolution of probabilistic weather forecasts with application in disease management. *Phytopathology*, *107*(2), 158–162. https://doi.org/10.1094/PHYTO-07-16-0256-R

Ishikawa, T., Barnston, A. G., Kastens, K. A., Louchouarn, P., & Ropelewski, C. F. (2005). Climate forecast maps as a communication and decision-support tool: An empirical test with prospective policy makers. *Cartography and Geographic Information Science*, *32*(1), 3–16. https://doi.org/10.1559/1523040053270747

Jolliffe, I. T., & Stephenson, D. B. (Eds.) (2012). *Forecast verification: A practitioner's guide in atmospheric science* (2nd ed.). Wiley-Blackwell.

Joslyn, S., Nadav-Greenberg, L., & Nichols, R. M. (2009). Probability of precipitation: Assessment and enhancement of end-user understanding. *Bulletin of the American Meteorological Society*, *90*(2), 185–194. https://doi.org/10.1175/2008BAMS2509.1

Knuth, K. H. (2013). Optimal data-based binning for histograms. *ArXiv:Physics/0605197*. Retrieved from http://arxiv.org/abs/physics/0605197

Łabędzki, L. (2017). Categorical forecast of precipitation anomaly using the standardized precipitation index SPI. *Water*, *9*(1), 8. https://doi.org/10.3390/w9010008

Loritz, R., Kleidon, A., Jackisch, C., Westhoff, M., Ehret, U., Gupta, H., & Zehe, E. (2019). A topographic index explaining hydrological similarity by accounting for the joint controls of runoff formation. *Hydrology and Earth System Sciences*, *23*(9), 3807–3821. https://doi.org/10.5194/hess-23-3807-2019

Mason, S. J., & Weigel, A. P. (2009). A generic forecast verification framework for administrative purposes. *Monthly Weather Review*, *137*(1), 331–349. https://doi.org/10.1175/2008MWR2553.1

Masoumi, F., & Kerachian, R. (2010). Optimal redesign of groundwater quality monitoring networks: A case study. *Environmental Monitoring and Assessment*, *161*(1–4), 247–257. https://doi.org/10.1007/s10661-008-0742-3

Mogheir, Y., de Lima, J. L. M. P., & Singh, V. P. (2003). Assessment of spatial structure of groundwater quality variables based on the entropy theory. *Hydrology and Earth System Sciences*, *7*(5), 707–721. https://doi.org/10.5194/hess-7-707-2003

Montgomery, D. C., & Runger, G. C. (2014). *Applied statistics and probability for engineers* (6th ed.). John Wiley and Sons, Inc.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*(2), 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

Murphy, A. H. (1998). The early history of probability forecasts: Some extensions and clarifications. *Weather and Forecasting*, *13*(1), 5–15. https://doi.org/10.1175/1520-0434(1998)013<0005:TEHOPF>2.0.CO;2

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*(7), 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2

Ning, Y., Ding, W., Liang, G., He, B., & Zhou, H. (2021). An analytical risk analysis method for reservoir flood control operation considering forecast information. *Water Resources Management*, *35*(7), 2079–2099. https://doi.org/10.1007/s11269-021-02795-6

North, R., Trueman, M., Mittermaier, M., & Rodwell, M. J. (2013). An assessment of the SEEPS and SEDI metrics for the verification of 6 h forecast precipitation accumulations: Assessment of SEEPS and SEDI for 6 h precipitation accumulations. *Meteorological Applications*, *20*(2), 164–175. https://doi.org/10.1002/met.1405

Pechlivanidis, I. G., Jackson, B., Mcmillan, H., & Gupta, H. V. (2016). Robust informational entropy-based descriptors of flow in catchment hydrology. *Hydrological Sciences Journal*, *61*(1), 1–18. https://doi.org/10.1080/02626667.2014.983516

Peng, Y., Xu, W., & Liu, B. (2017). Considering precipitation forecasts for real-time decision-making in hydropower operations. *International Journal of Water Resources Development*, *33*(6), 987–1002. https://doi.org/10.1080/07900627.2016.1219942

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1986). *Numerical recipes: The art of scientific computing* (1st ed.). Cambridge University Press.

Ridolfi, E., Montesarchio, V., Russo, F., & Napolitano, F. (2011). An entropy approach for evaluating the maximum information content achievable by an urban rainfall network. *Natural Hazards and Earth System Sciences*, *11*(7), 2075–2083. https://doi.org/10.5194/nhess-11-2075-2011

Ridolfi, E., Rianna, M., Trani, G., Alfonso, L., Di Baldassarre, G., Napolitano, F., & Russo, F. (2016). A new methodology to define homogeneous regions through an entropy based clustering method. *Advances in Water Resources*, *96*, 237–250. https://doi.org/10.1016/j.advwatres.2016.07.007

Rodwell, M. J., Richardson, D. S., Hewson, T. D., & Haiden, T. (2010). A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *136*(650), 1344–1363. https://doi.org/10.1002/qj.656

Ruddell, B. L., & Kumar, P. (2009). Ecohydrologic process networks: 1. Identification. *Water Resources Research*, *45*(3). https://doi.org/10.1029/2008WR007279

Särndal, C. E. (1974). A comparative study of association measures. *Psychometrika*, *39*(2), 165–187. https://doi.org/10.1007/BF02291467

Schirmer, M., & Jamieson, B. (2015). Verification of analysed and forecasted winter precipitation in complex terrain. *The Cryosphere*, *9*(2), 587–601. https://doi.org/10.5194/tc-9-587-2015

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610. https://doi.org/10.1093/biomet/66.3.605

Sharma, K., Ashrit, R., Kumar, S., Milton, S., Rajagopal, E. N., & Mitra, A. K. (2021). Unified model rainfall forecasts over India during 2007–2018: Evaluating extreme rains over hilly regions. *Journal of Earth System Science*, *130*(2), 82. https://doi.org/10.1007/s12040-021-01595-1

Shi, X., Wood, A. W., & Lettenmaier, D. P. (2008). How essential is hydrologic model calibration to seasonal streamflow forecasting? *Journal of Hydrometeorology*, *9*(6), 1350–1363. https://doi.org/10.1175/2008JHM1001.1

Sigaroodi, S. K., Chen, Q., Ebrahimi, S., Nazari, A., & Choobin, B. (2014). Long-term precipitation forecast for drought relief using atmospheric circulation factors: A study on the Maharloo Basin in Iran. *Hydrology and Earth System Sciences*, *18*(5), 1995–2006. https://doi.org/10.5194/hess-18-1995-2014

Singh, V. P. (2013). *Entropy theory and its application in environmental and water engineering*. Wiley-Blackwell.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, *21*(153), 65–66. https://doi.org/10.1080/01621459.1926.10502161

Su, X., Yuan, H., Zhu, Y., Luo, Y., & Wang, Y. (2014). Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. *Journal of Geophysical Research: Atmospheres*, *119*(12), 7292–7310. https://doi.org/10.1002/2014JD021733

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., et al. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, *97*(1), 49–67. https://doi.org/10.1175/BAMS-D-13-00191.1

Thiesen, S., Darscheid, P., & Ehret, U. (2019). Identifying rainfall-runoff events in discharge time series: A data-driven method based on information theory. *Hydrology and Earth System Sciences*, *23*(2), 1015–1034. https://doi.org/10.5194/hess-23-1015-2019

Thiesen, S., Vieira, D. M., Mälicke, M., Loritz, R., Wellmann, J. F., & Ehret, U. (2020). Histogram via entropy reduction (HER): An information-theoretic alternative for geostatistics. *Hydrology and Earth System Sciences*, *24*(9), 4523–4540. https://doi.org/10.5194/hess-24-4523-2020

Topp, C., Wang, W., Cloy, J., Rees, R., & Hughes, G. (2013). Information properties of boundary line models for N_2O emissions from agricultural soils. *Entropy*, *15*(3), 972–987. https://doi.org/10.3390/e15030972

Tovo, A., Formentin, M., Favretti, M., & Maritan, A. (2016). Application of optimal data-based binning method to spatial analysis of ecological datasets. *Spatial Statistics*, *16*, 137–151. https://doi.org/10.1016/j.spasta.2016.02.006

Wang, B., & Zhou, H. (2006). *Theory, method and application of reservoir dynamic control of the limited water level (in Chinese)*. China Water & Power Press.

Wang, H., Zhong, P., Zhu, F., Lu, Q., Ma, Y., & Xu, S. (2021). The adaptability of typical precipitation ensemble prediction systems in the Huaihe River basin, China. *Stochastic Environmental Research and Risk Assessment*, *35*(2), 515–529. https://doi.org/10.1007/s00477-020-01923-9

Wang, W., Wang, D., Singh, V. P., & Wang, Y. (2018). Spatial-temporal evaluation of rain-gauge network based on entropy theory. *EPiC Series in Engineering*, *3*, 2293–2284. https://doi.org/10.29007/1kc9

Weijs, S. V., Schoups, G., & van de Giesen, N. (2010). Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, *14*(12), 2545–2558. https://doi.org/10.5194/hess-14-2545-2010

Weijs, S. V., van Nooijen, R., & van de Giesen, N. (2010). Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, *138*(9), 3387–3399. https://doi.org/10.1175/2010MWR3229.1

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Elsevier.

Xi, S., Wang, B., Liang, G., Li, X., & Lou, L. (2010). Inter-basin water transfer-supply model and risk analysis with consideration of rainfall forecast information. *Science China Technological Sciences*, *53*(12), 3316–3323. https://doi.org/10.1007/s11431-010-4170-6

Xu, L., Chen, N., Zhang, X., & Chen, Z. (2020). A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. *Climate Dynamics*, *54*(7–8), 3355–3374. https://doi.org/10.1007/s00382-020-05173-x

Zhou, H., Tang, G., Li, N., Wang, F., Wang, Y., & Jian, D. (2011). Evaluation of precipitation forecasts from NOAA global forecast system in hydropower operation. *Journal of Hydroinformatics*, *13*(1), 81–95. https://doi.org/10.2166/hydro.2010.005