Geoscience Letters

RESEARCH LETTER

Open Access

Hypothesis testing for performance evaluation of probabilistic seasonal rainfall forecasts

Ke-Sheng Cheng[1*], Gwo-Hsing Yu[2], Yuan-Li Tai[2], Kuo-Chan Huang[2], Sheng-Fu Tsai[3], Dong-Hong Wu[4], Yun-Ching Lin[5], Ching-Teng Lee[5] and Tzu-Ting Lo[5]

Abstract

A hypothesis testing approach, based on the theorem of probability integral transformation and the Kolmogorov–Smirnov one-sample test, for performance evaluation of probabilistic seasonal rainfall forecasts is proposed in this study. By considering the probability distribution of monthly rainfalls, the approach transforms the tercile forecast probabilities into a forecast distribution and tests whether the observed data truly come from the forecast distribution. The proposed approach provides not only a quantitative measure for performance evaluation but also a cumulative probability plot for insightful interpretations of forecast characteristics such as overconfident, underconfident, mean-overestimated, and mean-underestimated. The approach has been applied for the performance evaluation of probabilistic season rainfall forecasts in northern Taiwan, and it was found that the forecast performance is seasonal dependent. Probabilistic seasonal rainfall forecasts of the Meiyu season are likely to be overconfident and mean-underestimated, while forecasts of the winter-to-spring season are overconfident. A relatively good forecast performance is observed for the summer season.

Keywords Probabilistic forecast, Seasonal rainfall, Performance evaluation, Hypothesis test

Introduction

Rainfall forecast plays an essential role in natural disaster prevention and mitigation. For such applications, very-short-range, short-range, to daily rainfall forecasts are needed. These forecasts can yield sub-hourly, hourly, and daily rainfall forecasts for the next several hours to days (Cuo et al. 2011; Shrestha et al. 2013; JMA 2018). Roberts et al. (2009) demonstrated the benefit of using high-resolution precipitation forecasts from numerical weather prediction (NWP) models for flood and short-term streamflow forecasting. Most NWP models are deterministic models. The uncertainty in the initial condition of weather variables; however, small, together with the model uncertainty, will lead to uncertainty in the forecast after a certain forecast lead time (Slingo and Palmer 2011). Hence, all NWP forecasts must be treated as probabilistic. Nowadays, accurate forecast of sub-hourly to daily rainfalls relies mainly on NWP models. However, machine learning techniques are also increasingly applied to short-range rainfall forecasts (Donlapark 2021; Chen and Wang 2022; Frnda et al. 2022).

In contrast to natural disaster prevention and mitigation, for which responsive actions are taken immediately before or after issuing the forecast, tasks like water

*Correspondence:
Ke-Sheng Cheng
rslab@ntu.edu.tw
[1] Center for Weather and Climate Disaster Research, National Taiwan University, Taipei, Taiwan, R.O.C.
[2] Taiwan Research Institute On Water Resources and Agriculture, New Taipei, Taiwan, R.O.C.
[3] Irrigation Agency, Ministry of Agriculture, Taipei, Taiwan, R.O.C.
[4] Ministry of Agriculture, Taiwan Agricultural Research Institute, Taichung, Taiwan, R.O.C.
[5] Central Weather Administration, Ministry of Transportation and Communications, Taipei, Taiwan, R.O.C.

resources planning and disaster management often need to make decisions several weeks or months in advance. For example, in a dry year, an irrigation manager needs to decide on paddy planting acreage and irrigation water allocation several months in advance (Tsai et al. 2023). Short-range rainfall forecasts cannot facilitate the data requirements for such long-term decision-making. Instead, information about the seasonal rainfall over the crop-growing season is crucial for such irrigation decision-making. Other examples of strategic planning for risk reduction using seasonal climate forecasts have also been documented (Dessai and Bruno Soares 2013; BoM and IFRC 2015) Nowadays, routine operational activities of global seasonal climate forecasts are being conducted by several meteorological forecast services, including the European Centre for Medium-Range Weather Forecasts, Japan Meteorological Agency, Met Office of UK, and the National Centers for Environmental Prediction of the United States.

Seasonal climate forecasts do not aim to forecast the day-to-day evolution of weather; instead, they provide estimates of seasonal-mean weather statistics over a region, typically up to 3 months ahead of the season in question (Weisheimer and Palmer 2014). In addition, weather models used to make seasonal forecasts are only approximate representations of reality. Thus, seasonal forecasts are probabilistic in nature, taking the form of occurrence probabilities over future events (Weisheimer and Palmer 2014). Probabilistic weather forecasting provides a range of plausible forecast results, which allows the forecaster to assess possible outcomes, and estimate the risks and probabilities of those outcomes. By considering perturbations to the initial conditions and stochastic parameterizations, ensemble forecasts are now fundamental to weather forecasting on all scales. It has been demonstrated that model-specific biases lead to under-dispersion in the ensemble; thus, the use of multi-model ensembles (MME) with greater reliability in the ensemble prediction system is pursued (Palmer et al. 2004; Slingo and Palmer 2011).

Probabilistic forecasts are probability statements about future outcomes; however, they are not necessarily issued as a probability for an event, such as the probability of raining or not raining. WMO (2020) recommended that operational seasonal forecasts be in a probabilistic format and that the probabilistic nature of seasonal forecasts be emphasized with a description of the probabilities used and their meaning. Different types of probabilistic seasonal forecasts can be issued (Troccoli et al. 2008). The most common type of probabilistic seasonal climate forecasts is to present the probabilities for the variable of interest, such as monthly rainfall or temperature, to fall into individual tercile categories. The tercile categories represent data ranges of below-normal, normal, and above-normal, and are determined based on the observed data within a specific historical period such as 1981 to 2010. Another type of probabilistic forecast is to present the probability density function or the cumulative distribution function of the forecast variable, conditioned on the current weather condition. This will give more complete and detailed information about the forecast variable; however, it may also be difficult to interpret for many end users.

Since the probabilistic forecasts do not yield specific values of the forecast variables, for example, rainfall amounts or temperatures, the forecast performance cannot be assessed using the attributes of forecast quality such as accuracy or correctness. In addition, forecasting skills can be evaluated only when a large number of similar forecasts are available. Many measures for performance evaluation of probabilistic forecasts exist in the literature (Bröcker and Smith 2007; Broecker 2012; Laio and Tamea 2007; Wilks 2019). All these measures are statistical characterization of the relationship between the observations and their corresponding forecasts. Two widely used measures are briefly described below.

Brier score (BS). Let y and o represent the probability forecast and the observation for probabilistic forecasting of an event E, respectively. The Brier score (Eq. 1) is defined as the mean squared error of the probability forecasts, considering that the observation is 1 if event E occurs and that the observation is 0 if event E does not occur:

$$BS = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - o_i\right)^2, 0 \le BS \le 1 \tag{1}$$

where n is the total number of (forecast y, observation o) pairs.

The forecast probabilities often only assume a few levels, such as multiples of 0.1. If there are k forecast probability levels, i.e., $y_i, i = 1, 2, \cdots, k$, then the above Brier score can be further decomposed into three terms (Murphy 1973; Troccoli et al. 2008; Wilks 2019):

$$BS = \sum_{i=1}^{k} p_i\left(y_i - \overline{o}_i\right)^2 - \sum_{i=1}^{k} p_i(\overline{o}_i - \overline{o})^2 + \overline{o}(1-\overline{o}) \tag{2a}$$

$$p_i = \frac{n_i}{n} \tag{2b}$$

where n_i is the number of forecasts given that probability level y_i was forecast, $\overline{o}_i$ is the average of all observations with corresponding forecast probability y_i, and $\overline{o}$ is the average of all observations, i.e., the occurrence probability of event E. The first decomposed term in Eq. (2a)

summarizes the conditional bias of the forecasts and is called the *reliability*.

For events with multi-category outcomes, as is the case of the tercile-category probabilistic forecast, the following multi-category Brier score can be calculated:

$$BS = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-o_{ij}\right)^2 = \sum_{j=1}^{m}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_{ij}-o_{ij}\right)^2\right] = \sum_{j=1}^{m} BS_j \quad (3)$$

where m is the number of outcome categories and BS_j is the Brier score for the event of category-j occurrence.

Brier scores close to zero indicate good forecast performance. However, there is no single standard for how small or large the Brier score should be for a model with good or poor forecast performance. For example, it is difficult to interpret the performance as good or bad for a forecast model with a Brier score of 0.35.

Reliability diagram. For a given binary event E, the reliability diagram is a graph that shows the correspondence of the forecast probabilities (y_i) with the observed relative frequency of occurrence ($\bar{o}_i$) of event E, given the forecast. The forecasts are considered reliable when the forecast probability is an accurate estimation of the relative frequency of the predicted outcome (Murphy 1993). The reliability diagram plots as a diagonal line for perfect forecasts, as illustrated in Fig. 1. Previous studies (Endris et al. 2021; Xu 2022) evaluated PSRF performance by considering regional or global probabilistic forecasts. In these studies, grid sizes of 0.5° and 1° were adopted for seasonal rainfall forecasts. Probabilistic forecasts at all grids within a specific region were combined to gain a large sample size, i.e. the number of PSRF runs, for the construction of reliability diagrams.

In a reliability diagram, forecast probabilities are grouped into a few probability levels, making each level have only a limited number of forecasts for calculation of its relative frequency ($\bar{o}_i$). Even the reliability diagrams of a perfectly reliable forecast system can exhibit deviations from the diagonal. Thus, evaluating a forecast system requires some idea as to how far the observed relative frequencies are expected to be from the diagonal if the forecast system is reliable (Bröcker and Smith 2007). Unlike the reliability term in Eq. (2a), which is a scalar summary measure, the reliability diagram uses k pairs of $(y_i, \bar{o}_i)$ to describe various properties, such as the overconfident, underconfident, well-calibrated, wet bias, and dry bias, of the probability forecasts (Wilks 2019; WMO 2020). However, it is difficult to quantitatively compare the forecast performance of different models using graphical diagnostic tools like the reliability diagram.

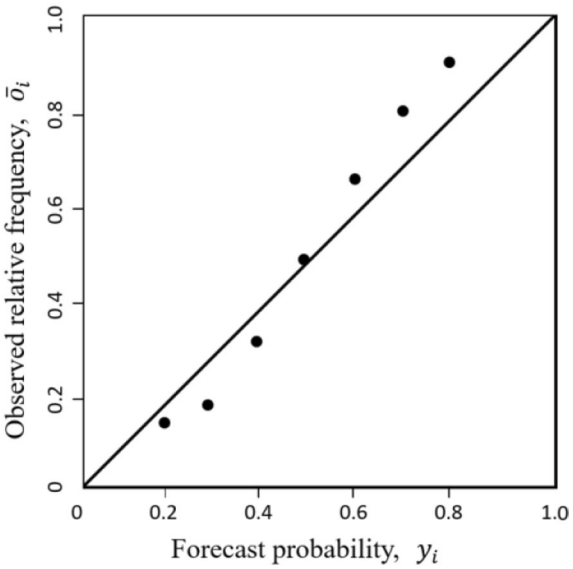

Fig. 1 Exemplar reliability diagram. Dots represent the (forecast probability, observed relative frequency) pairs of probability forecasts. The diagonal line represents perfect forecasts

In this paper, we focus on probabilistic seasonal rainfall forecast (PSRF), and hereinafter, monthly rainfalls are the climatological variable under investigation. Most PSRF systems consider tercile categories and yield tercile forecast probabilities, i.e., probabilities for monthly rainfalls of ℓ -month lead time to fall into individual tercile categories. Usually, probabilistic forecasts of 1-, 2-, and 3-month lead times are issued. Such practices require determining two tercile thresholds from monthly rainfall observations of a historical period. Each tercile category defines a dichotomous, or binary, event that monthly rainfalls will or will not fall into this tercile category. Let the below-normal, normal, and above-normal tercile categories be expressed by C_1, C_2, and C_3, respectively, and their corresponding events be E_1, E_2, and E_3. A forecast that yields $100p$% probability for C_1 can be interpreted as that there is a $100p$% chance that event E_1 will occur. Each forecast run results in three tercile forecast probabilities, or equivalently, the occurrence probabilities of E_1, E_2, and E_3. After a large number of forecast runs have been conducted, one can construct the reliability diagrams of events E_1, E_2, and E_3, respectively. However, when these reliability diagrams show different patterns, evaluating the overall performance of probabilistic forecasts may become complicated. Although the tercile thresholds of monthly rainfalls were calculated using historical observations, most PSRF systems do not consider the probability distribution properties of monthly rainfalls, including the distribution type and parameters. We believe

that considering the probability distribution of monthly rainfalls can lead to a more insightful evaluation of PSRF Systems.

In addition, a question that naturally arises when evaluating the performance of a PSRF system is whether the observed rainfalls truly come from the forecast distribution. This question can be dealt with by conducting statistical hypothesis tests, also known as the goodness-of-fit (GOF) tests. The Chi-squared test and the one-sample Kolmogorov–Smirnov (KS) test are the most widely used, particularly in the fields of water resources and hydrologic science (Kite 1977; Vlček and Huth 2009; Tarnavsky et al. 2012; Hamed and Rao 2019). Therefore, we propose a nonparametric goodness-of-fit test approach based on the Kolmogorov–Smirnov statistic for evaluating the performance of probabilistic seasonal forecasts.

This study aims to overcome the above difficulties in PSRF performance evaluation based on the Brier score and the reliability diagram. The proposed approach is statistically tractable and does not require using different reliability diagrams for below-normal, normal, and above-normal events or separating forecast probabilities into a few probability levels. Specifically, the main research goals of this study are to (1) provide a clear criterion for PSRF performance evaluation based on the KS hypothesis test and (2) derive a metric that does not need to separately evaluate the PSRF performance for the three tercile categories.

Methodology

In Taiwan, the Central Weather Administration (CWA) routinely issues probabilistic seasonal rainfall forecasts for the next 3 months at the end of the current month. Let X represent the monthly rainfalls of a specific month, say August, and q_1 and q_2 be the lower and upper tercile thresholds of X, respectively. Probabilistic rainfall forecasts for August can be issued at the end of May, June, and July, with 3-, 2-, and 1-month lead time, respectively. Let Y represent the forecast monthly rainfall of August under the current weather conditions. We shall refer to the cumulative distribution functions (CDF) of X and Y as the climate distribution and the forecast (or conditional) distribution, respectively. We further assume that X and Y are of the same distribution type with two parameters. A forecast run yields three forecast probabilities, say $\left(p_{E_1}, p_{E_2}, 1-p_{E_1}-p_{E_2}\right)$, where p_{E_1} and p_{E_2} are forecast probabilities of event E_1 (below-normal) and event E_2 (normal), respectively. We then have

$$F_Y(q_1;\alpha,\beta) = P(Y \le q_1) = p_{E_1} \tag{4a}$$

$$F_Y(q_2;\alpha,\beta) = P(Y \le q_2) = p_{E_1} + p_{E_2} \tag{4b}$$

where F_Y is the CDF of Y, and α and β are its parameters. Figure 2 illustrates the climate and forecast distributions and the cumulative probability of the observed rainfall, if the forecast distribution is true, of an exemplar forecast run.

For a two-parameter distribution, Cook (2010) showed how to solve for distribution parameters, given the two quantile conditions in Eqs. (4a) and (4b). If Y belongs to a location-scale family, its location (α) and scale (β) parameters can be obtained as follows:

$$\alpha = \frac{q_1 F_{Y^*}^{-1}\left(p_{E_1}+p_{E_2}\right) - q_2 F_{Y^*}^{-1}\left(p_{E_1}\right)}{F_{Y^*}^{-1}\left(p_{E_1}+p_{E_2}\right) - F_{Y^*}^{-1}\left(p_{E_1}\right)} \tag{5}$$

$$\beta = \frac{q_2 - q_1}{F_{Y^*}^{-1}\left(p_{E_1}+p_{E_2}\right) - F_{Y^*}^{-1}\left(p_{E_1}\right)} \tag{6}$$

where Y^* is the same location-scale family distribution with location and scale parameters being 0 and 1, respectively, and F_{Y^*} is the CDF of Y^*.

Assuming that forecast probabilities $\left(p_{E_1}, p_{E_2}, 1-p_{E_1}-p_{E_2}\right)$ of n forecast runs are available and let $o_i, i = 1, 2, \cdots, n$, be the corresponding monthly rainfall observations. If the probability distribution type of monthly rainfalls is known, the forecast distributions of individual forecast runs can be derived using Eqs. (5) and (6). By the theorem of probability integral transformation (PIT) (Mood et al. 1974), cumulative probabilities of o_i′s form a random sample of size n from the standard uniform distribution $U[0,1]$, if the observed rainfalls are truly from the forecast distribution F_Y, that is

$$u_i = F_Y(o_i;\alpha_i,\beta_i) \sim U[0,1], i = 1, 2, \cdots, n \tag{7}$$

where parameters (α_i, β_i) may vary among different forecast runs. The same concept has been applied to the PIT histogram and verification rank histogram to evaluate whether the forecast ensembles apparently include the observations being predicted as equiprobable members (Dawid 1984; Wilks 2019).

After the cumulative probabilities of the observed rainfalls have been calculated using Eq. (7), the one-sample KS GOF test can be conducted to test whether the observed monthly rainfalls truly come from the forecast distributions. This is equivalent to testing whether u_i′s are uniformly distributed. The KS statistic D_n is a measure of the maximum distance between the empirical CDF of the observed data and the CDF of the forecast, or hypothesized, distribution, that is

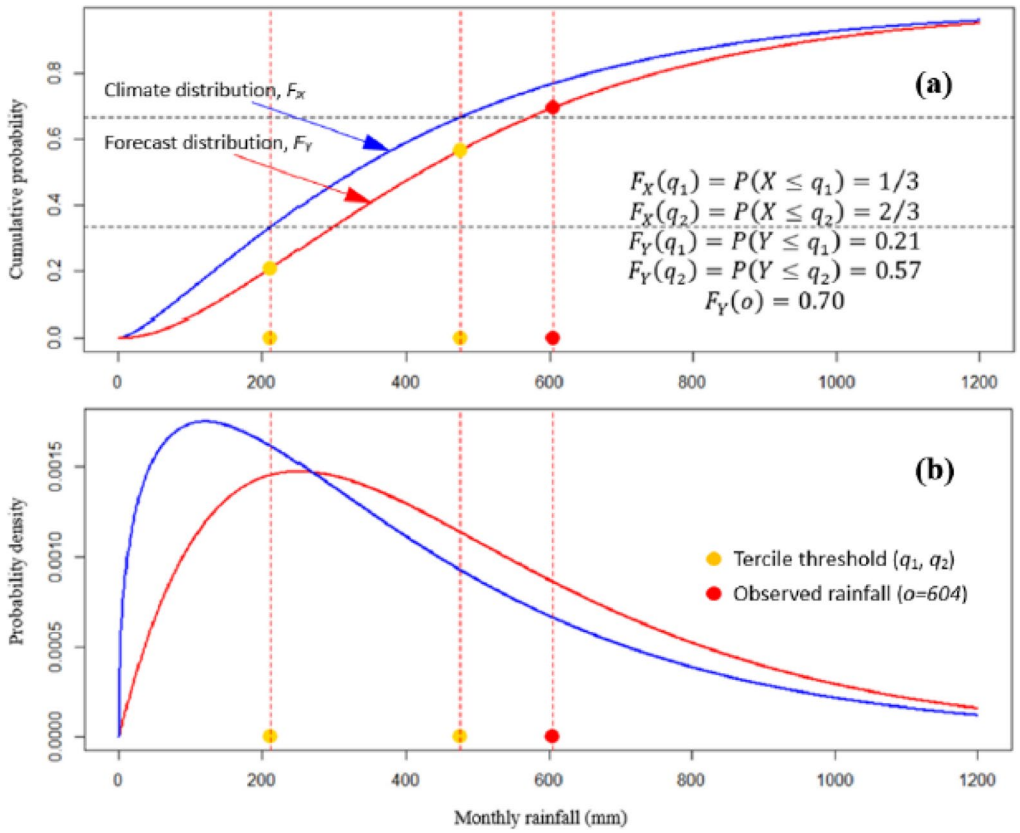

Fig. 2 Exemplar illustration of the climate and forecast distributions of monthly rainfall. **a** cumulative distribution functions; **b** probability density functions

$$D_n = \sup_{0 \le u \le 1} |F_n(u) - F_U(u)| \tag{8}$$

where F_n is the empirical CDF of u_i′s in Eq. (7) and F_U is the CDF of the standard uniform distribution. The critical region of the KS test statistic depends on the sample size n and is well-documented (Mood et al. 1974). If the KS test rejects the null hypothesis, it suggests that the forecast distribution does not properly characterize the observed data, or the observed data do not come from the forecast distribution.

Demonstration by stochastic simulation

To demonstrate the efficacy of the proposed approach, we conducted the following stochastic simulation to mimic the probabilistic forecasts and evaluate the forecast performance. Let W and X represent the monthly rainfalls of July and August, respectively, and q_1 and q_2 be the lower and upper tercile thresholds of X. We can think of X as the climate distribution of monthly rainfall of August, and W as the current weather condition that leads us to make a probabilistic forecast. In addition, let Y be the forecast monthly rainfall of August given the observed value of W, i.e., the conditional distribution of X given W. In our simulation, we assume that W and X form a bivariate normal distribution with the following parameters:

$$W \sim N(\mu_W = 860, \sigma_W = 279.28) \tag{9a}$$

$$X \sim N(\mu_X = 745, \sigma_X = 219.09) \tag{9b}$$

$$\rho_{WX} = 0.16 \tag{9c}$$

where μ, σ, and ρ represent the expected value, standard deviation, and correlation coefficient, respectively.

The above parameters were set for demonstration purposes by considering the long-term average monthly rainfalls of July and August for the Shihmen Reservoir watershed and Tsengwen Reservoir watershed, the two largest reservoirs in Taiwan (NCDR, n.d.; see Supplementary Information SI 1). Although these parameters are not exactly the same as the monthly rainfall statistics of the two reservoirs, they represent realistic amounts of monthly rainfall in summer in Taiwan. Figure 3 demonstrates a scatter plot of 10,000 sample pairs of (W, X)

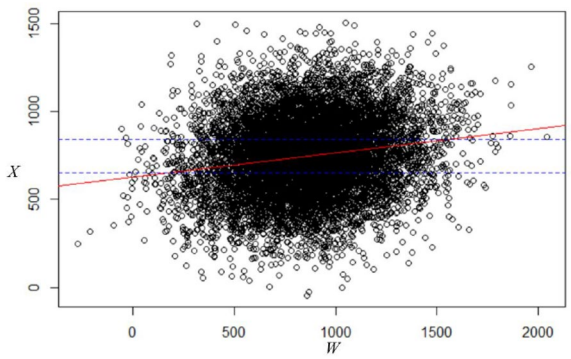

Fig. 3 Scatter plot of 10,000 sample pairs from a bivariate normal distribution. The red line represents the regression line. Blue dashed lines mark the tercile thresholds of *X*

from the above bivariate normal distribution. The lower and upper tercile thresholds of X are 650.63 and 839.37, respectively.

Given an observed monthly rainfall of July, say w, we expect the monthly rainfall of August to be from the following condition normal distribution:

$$f_Y(y) = f_{X|W}(y|W = w)$$

$$= \frac{1}{\sqrt{2\pi\left(1-\rho_{WX}^2\right)}\sigma_X}\exp\left\{-\frac{1}{2}\left[\frac{\left(y-\mu_X\right)-\rho_{WX}\frac{\sigma_X}{\sigma_W}(w-\mu_W)}{\sigma_X\sqrt{1-\rho_{WX}^2}}\right]^2\right\}. \quad (10a)$$

$$E(Y) = E(X|w) = \mu_X + \rho_{WX}\frac{\sigma_X}{\sigma_W}(w-\mu_W) \quad (10b)$$

$$Var(Y) = Var(X|w) = \sigma_X^2\left(1-\rho_{WX}^2\right) \quad (10c)$$

The above conditional distribution represents the forecast, or hypothesized, distribution, for monthly rainfall of August. For our stochastic simulation, a set of N random numbers of W, $\{w_i, i = 1, 2, \cdots, N\}$, were generated. This is equivalent to conducting N PSRF runs. For each w_i, the forecast distribution of monthly rainfall of August, i.e. $f_Y(y) = f_{X|W}(y|w_i)$, was determined using Eqs. (10b) and (10c).

Given an observed w_i, the observed monthly rainfalls of August, o_i, may or may not come from our forecast distribution. We assume that the true distribution of o_i is of the same distribution type as the forecast distribution, but with an inflated variance and/or increased mean value. The variance inflation factor (*VIF*) is defined as the ratio of the variance of the observed data to the variance of the forecast distribution. Similarly, the mean increase factor (*MIF*) is defined as the ratio of the expected value of the observed data to the expected value of the forecast distribution. If $VIF = MIF = 1$, the observed data are from the forecast distribution; otherwise, the forecast distribution does not correctly characterize the observed data. We then generated an observed value, say o_i, from the true distribution and calculated the cumulative probability $F_Y(o_i) = u_i$. The algorithm for stochastic simulation of PSRF performance evaluation using the KS test is illustrated in Fig. 4.

Suppose the probability distribution of the observed data and the forecast distribution differ only in their variances ($MIF=1$), the empirical CDF, $F_n(u)$, and the hypothesized CDF, $F_U(u)$, would exhibit patterns, as illustrated in Fig. 5 ($N=1000$) and Fig. 6 ($N=100$). Panel (a) in Fig. 5 shows that when the observed data are from the forecast distribution ($VIF=1$), $F_n(u)$ and $F_U(u)$ are nearly identical (well-calibrated), and the null hypothesis was not rejected at 5% level of significance ($p=0.690$). By contrast, panels (b) and (c) show rejection of the null hypothesis for underconfident ($VIF<1$) and overconfident forecasts ($VIF>1$), respectively. Although the corresponding reliability diagrams shown in panels (d), (e), and (f) seem to suggest a good correspondence between the forecast probability and the observed probability, they do not provide a quantitative measure of the forecast performance. When the sample size is reduced to 100, panels (a), (b), and (c) in Fig. 6 demonstrate similar patterns as in Fig. 5, but with larger deviations between $F_n(u)$ and $F_U(u)$. However, the reliability diagrams in panels (d), (e), and (f) of Fig. 6 show erratic patterns, making it difficult to evaluate the forecast performance. It is worthy to observe the $F_n(u) \sim F(u)$ patterns in Figs. 5 and 6. When $VIF < 1$, $F_n(u)$ falls below $F(u)$, with a concave form, in the lower tercile range and falls above $F(u)$, with a convex form, in the upper tercile range. Whereas when $VIF > 1$, $F_n(u)$ falls above $F(u)$, with a convex form, in the lower tercile range and falls below $F(u)$, with a concave form, in the upper tercile range.

If the probability distribution of the observed data and the forecast distribution differ only in their means ($VIF=1$), then $F_n(u)$ and $F(u)$ exhibit unique patterns, as illustrated in Fig. 7. When $MIF < 1$, $F_n(u)$ falls above $F(u)$ and has a convex form, whereas when $MIF > 1$, $F_n(u)$ falls below $F(u)$ and has a concave form.

Demonstration of PSRF performance evaluation using the KS test

1. Setting parameters of a bivariate normal distribution BVN(*W*,*X*)
   $W \sim N(\mu_W = 860, \sigma_W = 279.28)$
   $X \sim N(\mu_X = 745, \sigma_X = 219.09)$
   $\rho_{XY} = 0.16$
2. Set the number of PSRF runs, *N*.
3. Simulate a random sample of size *N* of *W*, i.e., $w_i, i = 1,2,\cdots,N$.
4. Set the value of *VIF*.
   If $VIF = 1$, the variance of the observed data is the same as the variance of the forecast distribution.
   IF $VIF > 1$, the variance of the observed data is larger than the variance of the forecast distribution.
   IF $VIF < 1$, the variance of the observed data is smaller than the variance of the forecast distribution.
   Set the value of *MIF*.
   If $MIF = 1$, the mean of the observed data is the same as the mean of the forecast distribution.
   IF $MIF > 1$, the mean of the observed data is higher than the mean of the forecast distribution.
   IF $MIF < 1$, the mean of the observed data is lower than the mean of the forecast distribution.
5. for (i in 1: *N*) {
   Calculate parameters of the probability distribution of the observed data given w_i,
   $E(Y) = E(X|w_i), \quad Var(Y) = Var(X|w_i),$
   $E(O) = E(Y) \times MIF, \ Var(O) = Var(Y) \times VIF$
   Simulate an observed value, o_i, using parameters $E(O)$ and $Var(O)$
   Calculate the cumulative probability of o_i of the forecast distribution, i.e., $u_i = F_Y(o_i)$.
   }
6. Conduct KS test for $\{u_i, i = 1,2,\cdots,N\}$, with $H_0: U(0,1)$, at level of significance $\alpha = 0.05$.

Fig. 4 Algorithm for stochastic simulation of PSRF performance evaluation using the KS test

The above unique $F_n(u) \sim F(u)$ patterns can provide valuable insights into the characteristics, underconfident, overconfident, mean-underestimated (dry-biased), and mean-overestimated (wet-biased), of the PSRF results. For example, Fig. 8 demonstrates $F_n(u) \sim F(u)$ patterns for four (*VIF*, *MIF*) combinations. These patterns can be easily explained by the above insightful observations and can serve as guidelines to uncover the causes of PSRF results.

For a hypothesis test, the power of the test represents the probability of rejecting the null hypothesis when it is wrong. In the context of PSRF, if the null hypothesis is rejected, it suggests that the observed data are not from the forecast distribution. Thus, the power of the KS test represents the capability of invalidating a PSRF system when its tercile forecast probabilities fail to characterize the probability distribution of the observed data. To demonstrate the power function of the KS test under

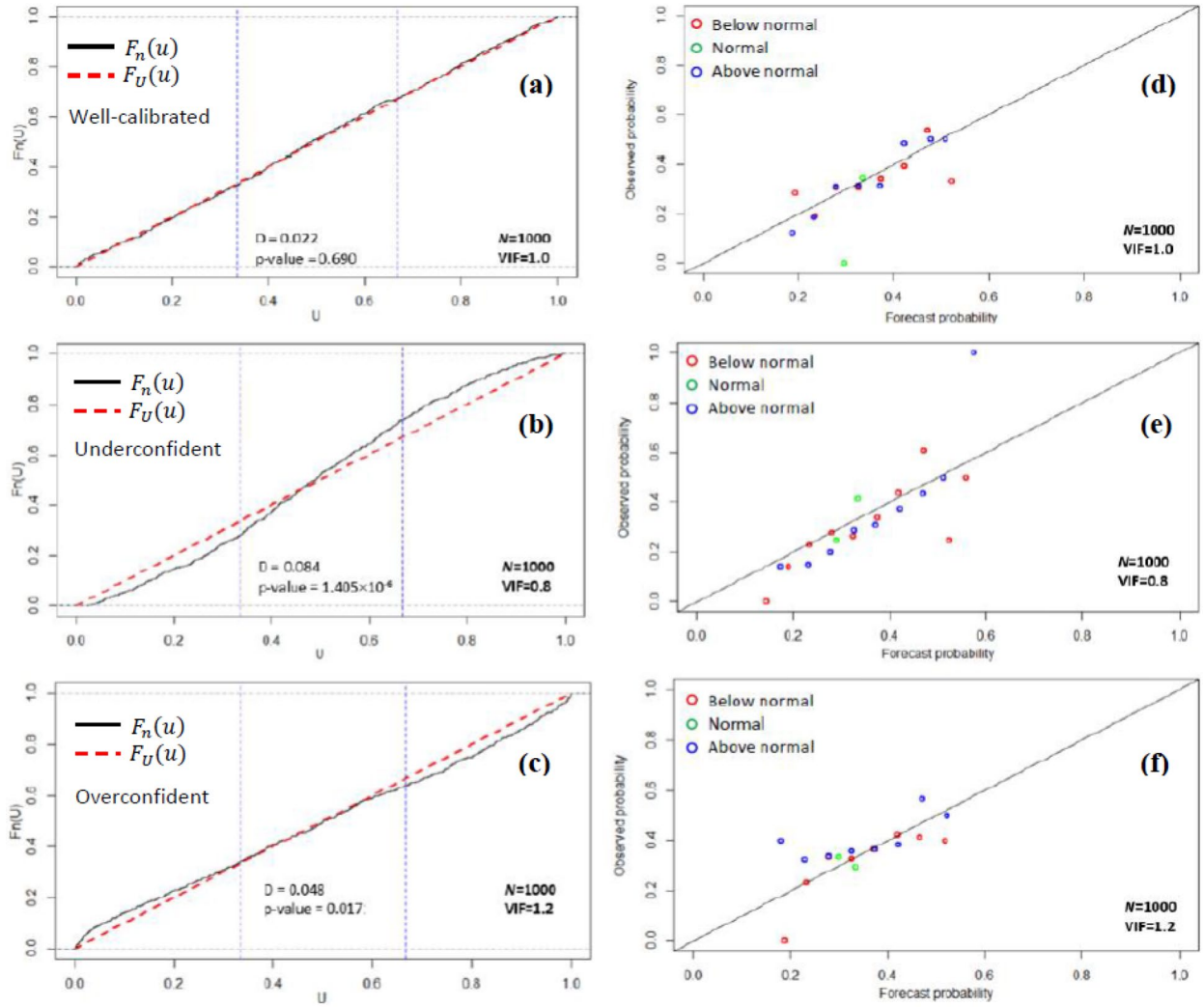

Fig. 5 Exemplar results of the KS test (left column) and the corresponding reliability diagrams (right column), 1000 PSRF runs. *D*: sample value of the KS statistic

different situations, we carried out 1,000 repeats of the simulation process in Fig. 4 for every selected combination of *VIF* (0.2–2.0 at increments of 0.1), *MIF* (0.9–1.1 at increments of 0.1), and *N* (100, 200–1000 at increments of 200) values. For a specific (*VIF*, *MIF*, *N*) combination, the power of the KS test is calculated as the proportion of the 1,000 repeats that rejected the null hypothesis. Figure 9 shows levelplots of the power of the KS test based on our stochastic simulation. Generally speaking, the power increases with the number of PSRF runs, and the *MIF* appears to have a higher effect on the power than does the *VIF*. Figure 10 shows the power function of the KS test when only the variation in variance ($MIF=1$) or mean ($VIF=1$) is considered. For $N=100$, the power function reaches 0.4 when the *VIF* is near 0.5 or 1.9, i.e., the variance of the observed data is 40% lower or higher than the variance of the forecast distribution. Whereas the same power level is reached when the *MIF* is 0.94 or 1.06, i.e., the mean of the observed data is 6% lower or higher than the mean of the forecast distribution. These results reveal that PSRF systems that overestimate/underestimate the mean are more likely to be invalidated by the KS test than those that overestimate/underestimate the variance.

Study case—performance evaluation for PSRF in northern Taiwan

At the end of a month, CWA issues probabilistic seasonal rainfall forecasts for four regions (North, Center, South, and East) in Taiwan, by considering the observed weather

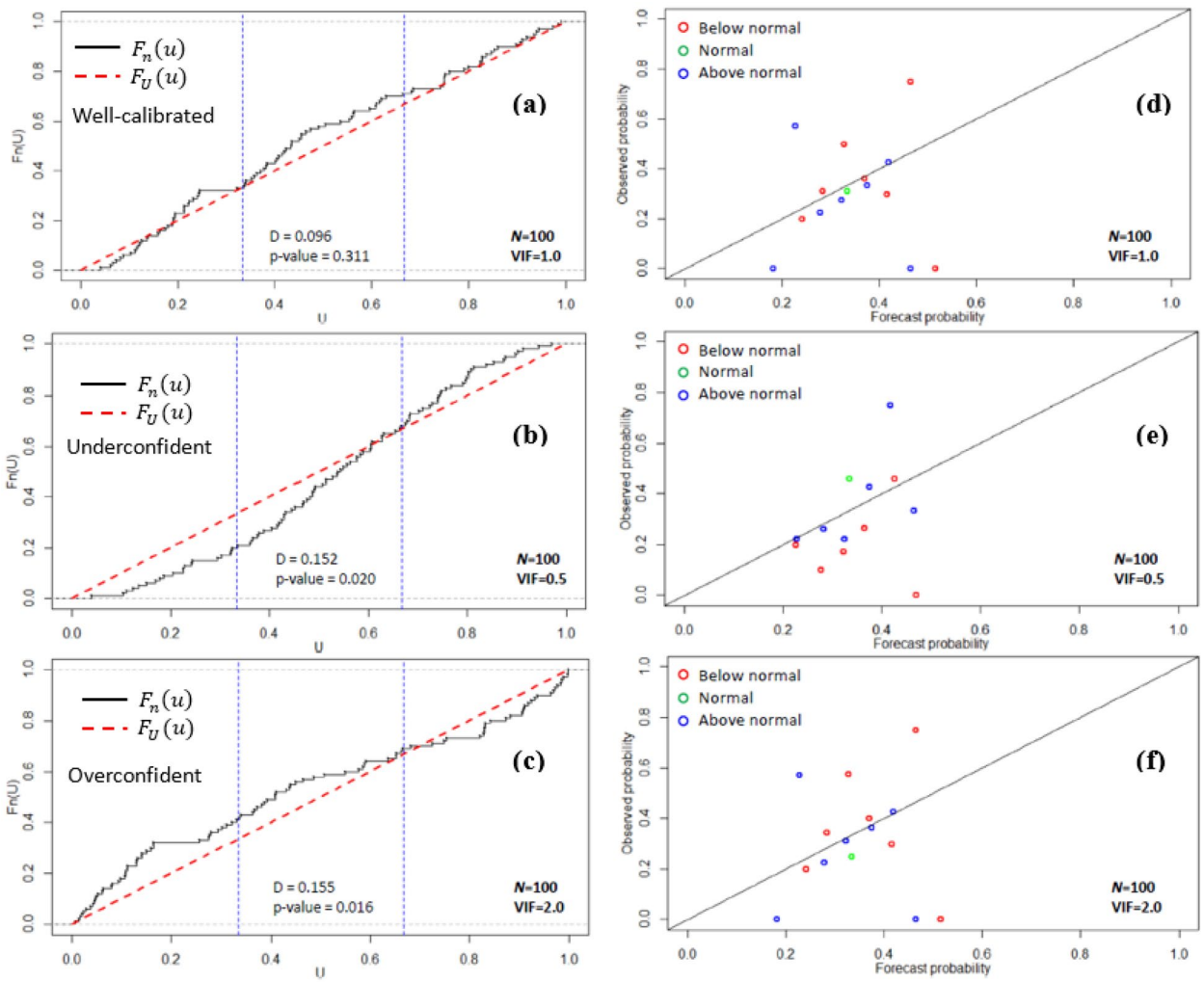

Fig. 6 Exemplar results of the KS test (left column) and the corresponding reliability diagrams (right column), 100 PSRF runs. *D*: sample value of the KS statistic

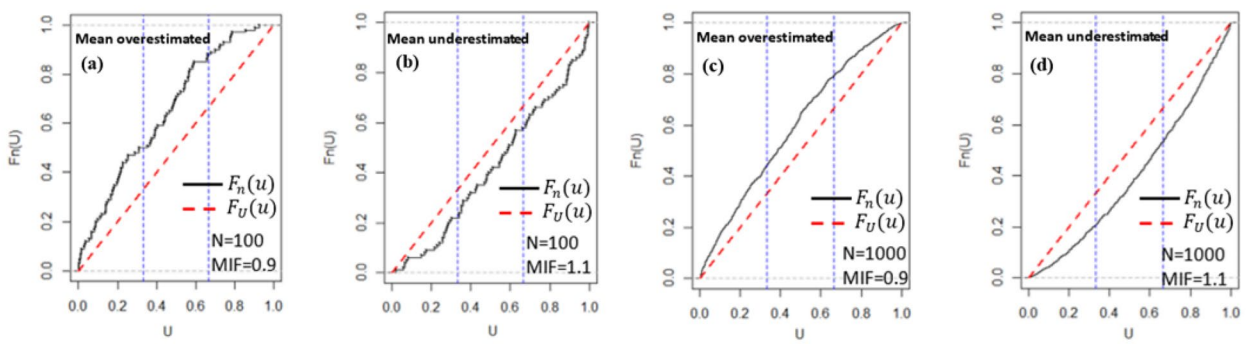

Fig. 7 $F_n(u) \sim F(u)$ patterns for changes in the mean of the forecast distribution

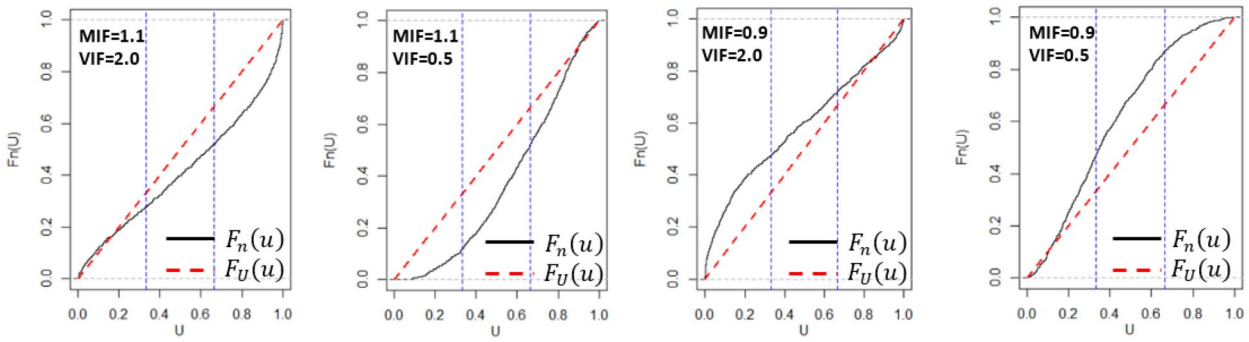

Fig. 8 $F_n(u) \sim F(u)$ patterns for various combinations of (*MIF*, *VIF*). Number of forecast runs $N = 1000$

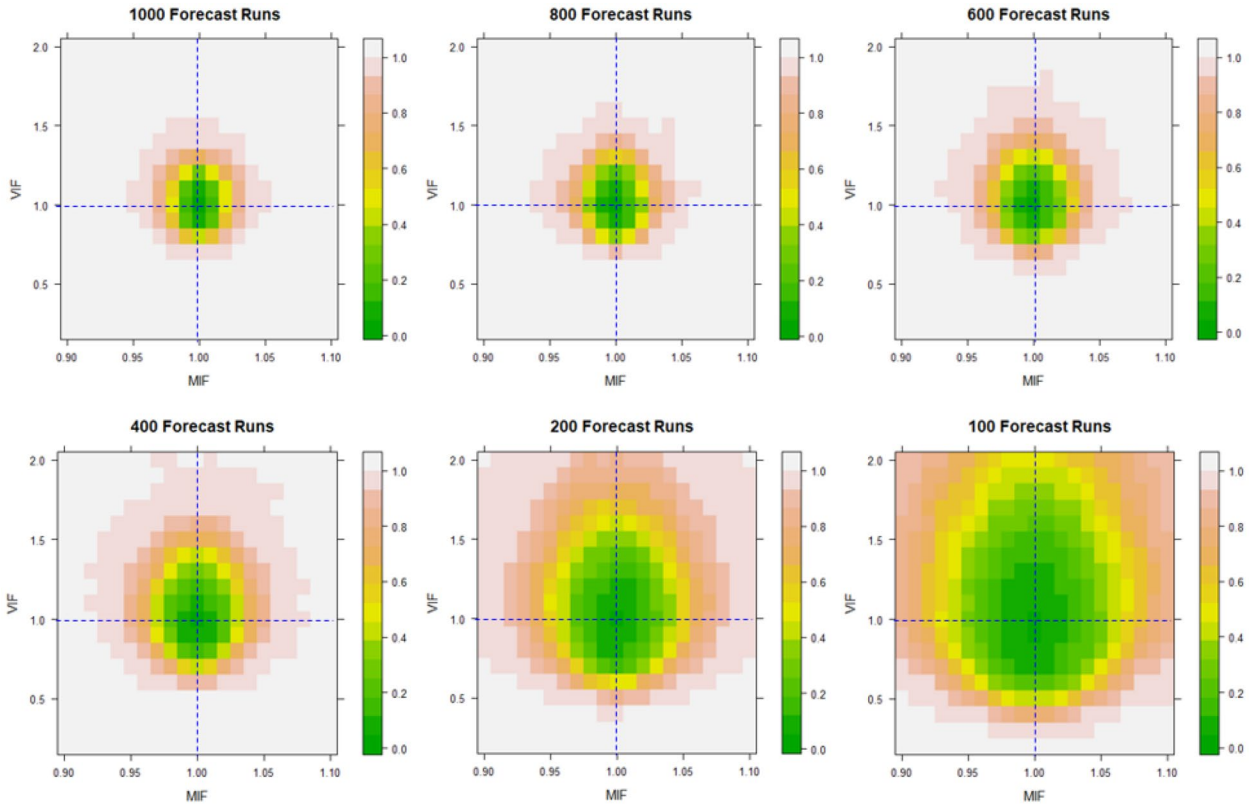

Fig. 9 Levelplots of the power of the KS test, with respect to various sample sizes, based on the stochastic simulation described in the third section. Power functions of the two dashed profiles (*MIF* = 1 and *VIF* = 1) are shown in Fig. 10

conditions and multi-model ensemble forecasts at a representative rainfall station in each region. Historical monthly rainfalls (1981–2020) and tercile forecast probabilities (2004–2020) for the North region were used in this study for PSRF performance evaluation. CWA calculated tercile thresholds (q_1, q_2) of individual months using 30 years of monthly rainfall observations at the representative Taipei station. These threshold values are updated every 10 years. For PSRF of 2001–2010, tercile thresholds were calculated using monthly rainfalls over the 1971–2000 period, whereas, for PSRF of 2011–2020, tercile thresholds were calculated using monthly rainfalls over the 1981–2010 period (see details in Supplementary Information SI 2).

A two-parameter distribution must be adopted to determine the forecast distribution of monthly rainfalls based on the tercile forecast probabilities issued by CWA. From the results of GOF tests for monthly rainfalls

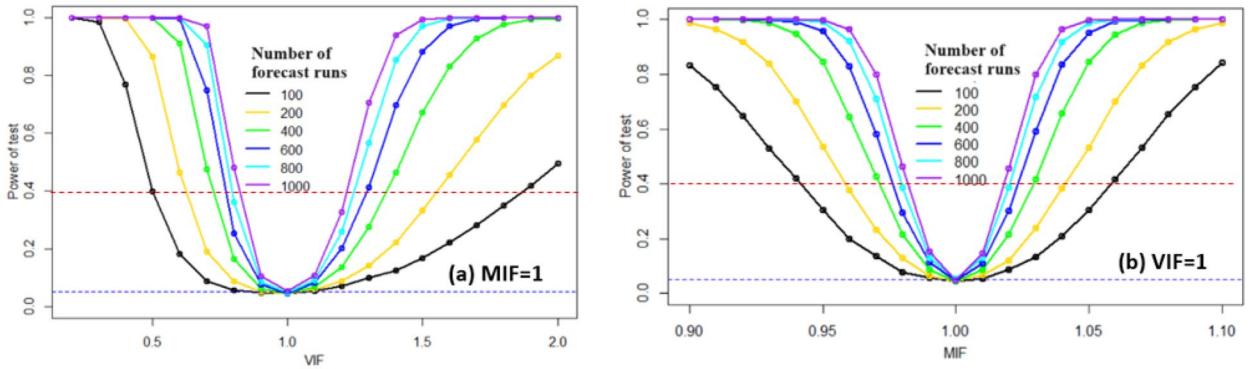

Fig. 10 Power functions of the KS test with respect to various sample sizes. **a** *MIF* = 1, **b** *VIF* = 1

(1981–2020) at Taipei station using *L*-moment-ratio diagrams (Liou et al. 2008; Wu et al. 2012), the following two-parameter log-normal distribution was chosen to fit the monthly rainfalls of individual months at Taipei station:

$$f_X(x) = \frac{1}{x\sqrt{2\pi}\sigma_{lnx}} e^{-\frac{1}{2}\left(\frac{lnx-\mu_{lnx}}{\sigma_{lnx}}\right)^2}, 0 < x < +\infty \quad (11)$$

where μ_{lnx} and σ_{lnx} are the expected value and standard deviation of lnX, respectively. Given the tercile thresholds (q_1, q_2) and tercile forecast probabilities (p_1, p_2) of a specific month, the location parameter (μ_{lnx}) and scale parameter (σ_{lnx}) of the log-normal forecast distribution could be determined using Eqs. (5) and (6). Cumulative probabilities of observed monthly rainfalls [see Eq. (7)] over the 2004–2020 period for Taipei station were then calculated using the corresponding forecast distributions.

Taiwan experiences heavy rainfalls caused by mesoscale convective systems, which is known as the Meiyu frontal rainfalls, in late spring or early summer (May–June), and by typhoons and convective storms in summer or early fall (July–October). Northeasterly monsoon also causes winter-to-spring (November to April) frontal rainfalls over the northeastern part of Taiwan. These prevalent storms differ in terms of their annual occurrence frequency, storm duration, and rainfall intensity (Cheng et al. 2024). Therefore, monthly rainfalls and the corresponding tercile forecast probabilities were partitioned into three groups, namely, the Meiyu season, summer season, and winter-to-spring season, and their PSRF performance evaluations were conducted separately.

Table 1 summarizes the results of the KS test for PSRF performance evaluation in northern Taiwan. For PSRF of the winter-to-spring season, the null hypothesis was rejected at 5% level of significance. For PSRF of the Meiyu season, the KS test rejected the null hypothesis at 10% level of significance. The higher level of significance was chosen for KS test of the Meiyu season for two reasons (Labovitz 1968; Kim and Choi 2021): (1) the smaller sample size (34 forecast runs) for the Meiyu season and (2) the size of the true difference between the means of the observed data and the hypothesized distribution is expected to be small for PSRF. For PSRF of the summer season, the null hypothesis was not rejected at 5% level of significance. If the KS test rejects the null hypothesis, it is likely that the observed data do not come from the forecast distribution, as has been explained in the Methodology section. The causes for rejecting the null hypothesis were further investigated by examining the $F_n(u) \sim F(u)$ patterns of PSRF of different seasons.

Table 1 Results of the KS test for PSRF performance evaluation in northern Taiwan

Forecast Lead (months)	Season	KS statistic, D_n	Sample size, n	p value
1	Meiyu	0.2191	34	0.0650**
	Summer	0.1211	68	0.2509
	Winter-to-spring	0.1620	102	0.0095*
2	Meiyu	0.2284	34	0.0484*
	Summer	0.1241	68	0.2263
	Winter-to-spring	0.1522	102	0.0178*
3	Meiyu	0.2154	34	0.0729**
	Summer	0.1082	68	0.3761
	Winter-to-spring	0.1553	102	0.0146*

* Significant at 5% level of significance

** Significant at 10% level of significance

Figure 11 illustrates the $F_n(u) \sim F(u)$ patterns of 1-, 2-, and 3-month lead PSRF at Taipei station for the Meiyu, summer, and winter-to-spring seasons. The $F_n(u) \sim F(u)$ patterns of these three groups are markedly different.

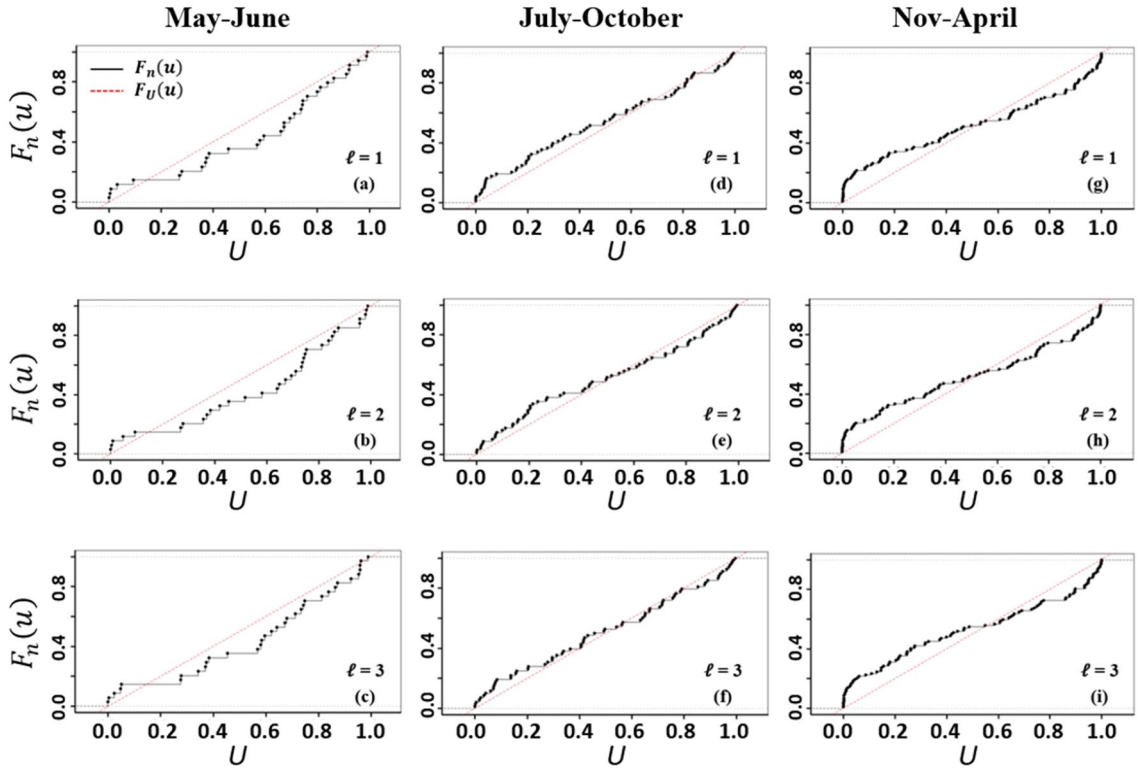

Fig. 11 $F_n(u) \sim F(u)$ patterns of ℓ-month lead PSRF at Taipei station

By referencing to Fig. 8, the PSRF of the Meiyu season is likely to be overconfident and mean-underestimated, while the PSRF of the winter-to-spring season is overconfident. A relatively good PSRF performance is observed for the summer season, with a minor degree of being overconfident and mean-overestimated. These results suggest that the performance of CWA's PSRF is seasonal-dependent. However, given the seasonal effect, the forecast lead time does not seem to affect the PSRF performance, as seen from the very similar empirical CDFs of different lead forecasts in Fig. 11.

Table 2 shows the multi-category Brier scores and reliabilities of PSRF in northern Taiwan. Generally speaking, the multi-category Brier scores and reliabilities of the Meiyu season are higher than those of the summer and winter-to-spring seasons, indicating poorer performance of the Meiyu season than other seasons. Such results are consistent with the evaluation by the KS test, although the Brier scores are less informative.

Reliability diagrams for PSRF of the Meiyu, summer, and winter-to-spring seasons are shown in Fig. 12. The reliability diagram of the Meiyu season appears to be more widely scattered away from the diagonal than other seasons. There are only a few forecast probability levels for each category. Notably, PSRF of the normal category (event E_2 in the Introduction section) has only 3 forecast

Table 2 Multi-category Brier scores and reliabilities of the PSRF in norther Taiwan

Brier score	**1-month lead**	**2-month lead**	**3-month lead**
Meiyu	0.6953	0.7146	0.7129
Summer	0.6773	0.6836	0.6896
Winter-to-spring	0.6965	0.6827	0.6857
Reliability	**1-month lead**	**2-month lead**	**3-month lead**
Meiyu	0.1016	0.0928	0.1154
Summer	0.0393	0.0471	0.0462
Winter-to-spring	0.0584	0.0434	0.0349

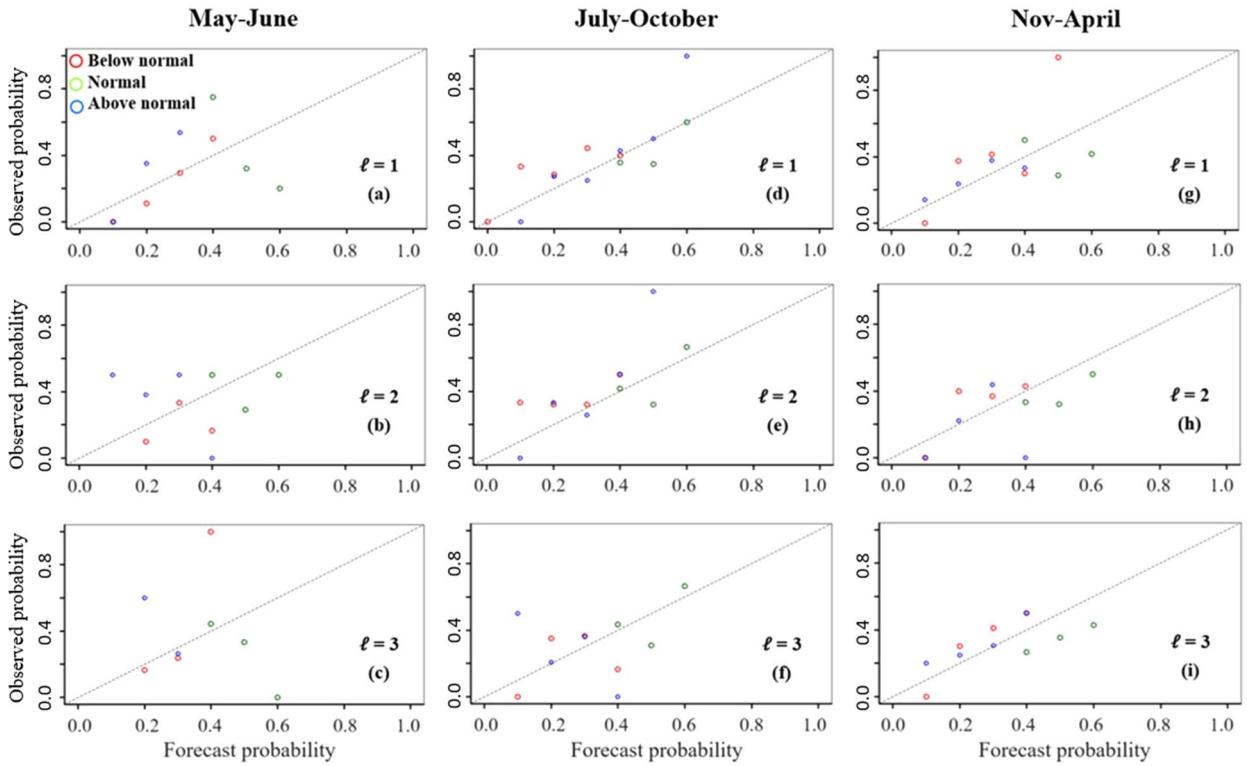

Fig. 12 Reliability plots for PSRF of different seasons in northern Taiwan. meiyu season: (**a**–**c**), summer season: (**d**–**f**), winter-to-spring season: (**g**–**i**)

probability levels, 40%, 50%, and 60%, regardless of the seasons and lead time. With a limited number of forecast probability levels, it is rather difficult to use the reliability diagrams shown in Fig. 12 to describe various properties, such as the overconfident, underconfident, well-calibrated, mean-overestimated, and mean-underestimated, of the forecast probabilities.

Table 3 further summarizes the frequencies of individual forecast probability levels with respect to different categories and seasons. The normal category was always forecast as having either a 40%, 50%, or 60% chance of occurrence. The 50% chance of the normal category occurrence accounts for 72% (79/110), 69% (151/220), and 61% (200/330) of the Meiyu, summer, and winter-to-spring events, respectively. Both the below-normal and above-normal categories were mostly forecast to have a 20–30% chance of occurrence. The 20–30% chance of the below-normal category occurrence accounts for 85% (94/110), 85% (188/220), and 80% (265/330) of the Meiyu, summer, and winter-to-spring events, respectively. The

Table 3 Frequency table of tercile forecast probabilities for different seasons

Season	Tercile category	Tercile forecast probability (%)							Average probability
		0	10	20	30	40	50	60	
Meiyu	Below normal		4	32	62	12			0.27
	Normal					22	79	9	0.49
	Above normal		3	64	42	1			0.24
Summer	Below normal	1	14	94	94	17			0.25
	Normal					49	151	20	0.49
	Above normal		8	90	106	10	5	1	0.26
Winter-to-spring	Below normal		7	85	180	57	1		0.29
	Normal					93	200	37	0.48
	Above normal		19	203	101	7			0.23

20–30% chance of the above-normal category occurrence accounts for 96% (106/110), 89% (196/220), and 92% (304/330) of the Meiyu, summer, and winter-to-spring events, respectively. Apparently, too many historical events were forecast to have a very high chance (50% and 60%) of normal category occurrence. Average forecast probabilities of the below-normal, normal, and above-normal categories for Meiyu, summer, and winter-to-spring events are also shown in Table 3. The average forecast probability of the normal category is higher than 48% for all seasons, while the average forecast probabilities of the below-normal and above-normal categories vary between 23% and 29%. Compared with the 33.3% occurrence probability for the tercile categories under the climate condition, the above results indicate overconfident forecasts for PSRF of all seasons in northern Taiwan.

Summary and conclusions

This study proposed a hypothesis testing approach to the performance evaluation of probabilistic seasonal rainfall forecasts. The approach first transforms the tercile forecast probabilities to a forecast distribution of monthly rainfalls, and, through the theorem of probability integral transformation, it enables the Kolmogorov–Smirnov hypothesis test of whether the observed monthly rainfalls truly come from the forecast distribution. Compared to other measures of PSRF performance evaluation, such as the Brier scores and reliability diagram, the proposed approach offers not only a quantitative measure but also insightful $F_n(u) \sim F(u)$ patterns to uncover the causes of the PSRF performance. Unlike the reliability diagrams, the $F_n(u) \sim F(u)$ patterns established by our approach do not need to separate the below-normal, normal, and above-normal events and 0.1-multiples forecast probability categories. The proposed approach has been applied to the performance evaluation of PSRF in northern Taiwan, and the following conclusions can be drawn from its results.

(1) CWA's PSRF performance is seasonal dependent. PSRF of the Meiyu season is likely to be overconfident and mean-underestimated, while PSRF of the winter-to-spring season is overconfident. A relatively good PSRF performance is observed for the summer season, with a minor degree of being overconfident and mean-overestimated.
(2) Given the seasonal effect, the forecast lead time does not affect the PSRF performance.
(3) The multi-category Brier scores and the frequency table of tercile forecast probabilities also indicate overconfident forecasts for PSRF of all seasons in northern Taiwan, supporting the findings of the proposed Kolmogorov–Smirnov hypothesis testing approach.

Abbreviations

CWA	Central Weather Administration
GOF	Goodness-of-fit
MME	Multi-model ensemble
NWP	Numerical weather prediction
PSRF	Probabilistic seasonal rainfall forecast

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40562-024-00341-x.

Additional file 1.

Acknowledgements

We acknowledge the funding support of the National Science and Technology Council (NSTC-112-2101-01-30-09) and the Irrigation Agency, Ministry of Agriculture, Taiwan, R.O.C.

Author contributions

KSC: conceptualization, formal analysis, methodology, supervision, writing. GHY: conceptualization, funding acquisition, resources. YLT: formal analysis, data curation, software, validation. KCH: data curation, software. SFT: conceptualization, funding acquisition. DHW: conceptualization, funding acquisition. YCL: methodology, data curation, validation. CTL: methodology, data curation, validation. TTL: methodology, data curation, validation.

Funding

This study received funding supports of the National Science and Technology Council (NSTC-112-2101-01-30-09) and the Irrigation Agency, Ministry of Agriculture, Taiwan, R.O.C.

Availability of data and materials

Data will be made available on request.

Declarations

Competing interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: KSC reports financial supports were provided by National Science and Technology Council (NSTC-112-2101-01-30-09) and Irrigation Agency, Ministry of Agriculture, Taiwan, R.O.C.

Received: 8 February 2024 Accepted: 8 May 2024
Published online: 27 May 2024

References

BoM and IFRC (2015) Linking seasonal forecasts with disaster preparedness in the Pacific: from information to action. Bureau of Meteorology, Australia Government and International Federation of Red Cross and Red Crescent Societies. http://www.climatecentre.org/downloads/files/IFRCGeneva/Seasonal%20Rainfall%20Watch%20Case%20Study%20FINAL.PDF Accessed 2 Nov 2023.

Bröcker J, Smith LA (2007) Increasing the reliability of reliability diagrams. Weather Forecast 22:651–661. https://doi.org/10.1175/WAF993.1

Broecker J (2012) Probability forecast. In: Jolliffe IT, Stephenson DB (eds) Forecast verification: a practitioner's guide in atmospheric science. John Wiley & Sons Ltd, Hoboken, pp 119–139

Chen G, Wang W-C (2022) Short-term precipitation prediction for contiguous United States using deep learning. Geophys Res Lett 49:e2022GL097904. https://doi.org/10.1029/2022GL097904
Cheng KS, Chen BY, Lin TW, Nakamura K, Ruangrassamee P, Chikamori H (2024) Rainfall frequency analysis using event-maximum rainfalls—an event-based mixture distribution modeling approach. Weather Clim Extremes 43:100634. https://doi.org/10.1016/j.wace.2023.100634
Cook J (2010) Determining distribution parameters from quantiles. UT MD Anderson Cancer Center Department of Biostatistics, Working Paper Series. https://www.johndcook.com/quantiles_parameters.pdf. Accessed 7 Nov 2023.
Cuo L, Pagano TC, Wang QJ (2011) A Review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. J Hydrometeorol 12:713–728. https://doi.org/10.2307/24912965
Dawid AP (1984) Present position and potential developments: some personal views: statistical theory: the prequential approach. J R Stat Soc Ser A 147:278–292
Dessai S, Bruno Soares M (2013) Literature review of the use of seasonal-to-decadal (S2D) predictions across all sectors. Deliverable report 12.1 of the EUPORIAS. https://euporias-test2.wdfiles.com/local--files/events-meetings/D12.1.pdf. Accessed 2 Nov 2023.
Donlapark P (2021) Short-term daily precipitation forecasting with seasonally-integrated autoencoder. Appl Soft Comput 102:107083. https://doi.org/10.1016/j.asoc.2021.107083
Endris HS, Hirons L, Segele ZT, Gudoshava M, Woolnough S, Artan GA (2021) Evaluation of the skill of monthly precipitation forecasts from global prediction systems over the Greater Horn of Africa. Weather Forecas 36:1275–1298. https://doi.org/10.1175/WAF-D-20-0177.1
Frnda J, Durica M, Rozhon J et al (2022) ECMWF short-term prediction accuracy improvement by deep learning. Sci Rep 12:7898. https://doi.org/10.1038/s41598-022-11936-9
Hamed K, Rao AR (2019) Flood frequency analysis (new directions in civil engineering). CRC Press, Boca Raton
JMA (2018) Very-short-range forecasts of precipitation. Japan Meteorological Agency. https://www.jma.go.jp/jma/en/Activities/qmws_2018/Presentation/3.1/Very-short-range%20Forecast%20of%20Precipitation.pdf. Accessed 31 Oct 2023
Kim JH, Choi I (2021) Choosing the level of significance: a decision-theoretic approach. Abacus 57:27–71. https://doi.org/10.1111/abac.12172
Kite GW (1977) Frequency and risk analysis in hydrology. Water Resources Publications, Littleton
Labovitz S (1968) Criteria for selecting a significance level: a note on the sacredness of. 05. Am Sociol 3:220–222
Laio F, Tamea S (2007) Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol Earth Syst Sci 11(1267–1277):2007. https://doi.org/10.5194/hess-11-1267-2007
Liou JJ, Wu YC, Cheng KS (2008) Establishing acceptance regions for L-moments based goodness-of-fit tests by stochastic simulation. J Hydrol 355:49–62
Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics. McGraw-Hill, New York
Murphy AH (1973) A new vector partition of the probability score. J Appl Meteor 12:595–600
Murphy AH (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecast 8:281–293
NCDR (n.d.) Weather and climate monitoring—average monthly rainfall. National Science and Technology Center for Disaster Reduction. https://watch.ncdr.nat.gov.tw/watch_monthlyrain. Accessed 6 Apr 2024
Palmer TN et al (2004) Development of a European multi-model ensemble system for seasonal-to-interannual prediction (DEMETER). Bull Am Meteorol Soc 85:853–872. https://doi.org/10.1175/BAMS-85-6-853
Roberts NM, Cole SJ, Forbes RM, Moore RJ, Boswell D (2009) Use of high-resolution NWP rainfall and river flow forecasts for advance warning of the Carlisle flood, North-West England. Meteorol Appl 16:23–44
Shrestha DL, Robertson DE, Wang QJ, Pagano TC, Hapuarachchi HAP (2013) Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. Hydrol Earth Syst Sci 17:1913–1931. https://doi.org/10.5194/hess-17-1913-2013
Slingo J, Palmer T (2011) Uncertainty in weather and climate prediction. Phil Trans R Soc A 369:4751–4767. https://doi.org/10.1098/rsta.2011.0161
Tarnavsky E, Mulligan M, Husak G (2012) Spatial disaggregation and intensity correction of TRMM-based rainfall time series for hydrological applications in dryland catchments. Hydrol Sci J 57(2):248–264
Troccoli A et al (2008) Seasonal climate: forecasting and managing risk. Springer Science + Business Media B.V., Dordrecht
Tsai SF, Wu DH, Yu GH, Cheng KS (2023) Risk-based irrigation decision-making for the Shihmen Reservoir Irrigation District of Taiwan. Paddy Water Environ 21:497–508. https://doi.org/10.1007/s10333-023-00943-9
Vlček O, Huth R (2009) Is daily precipitation Gamma-distributed?: Adverse effects of an incorrect use of the Kolmogorov-Smirnov test. Atmos Res 93(4):759–766. https://doi.org/10.1016/j.atmosres.2009.03.005
Weisheimer A, Palmer TN (2014) On the reliability of seasonal climate forecasts. J R Soc Interface 11:20131162. https://doi.org/10.1098/rsif.2013.1162
Wilks DS (2019) Statistical methods in the atmospheric sciences, 4th edn. Elsevier, Amsterdam
WMO (2020) Guidance on operational practices for objective seasonal forecasting. World Meteorological Organization, Geneva
Wu YC, Liou JJ, Su YF, Cheng KS (2012) Establishing acceptance regions for L-moments based goodness-of-fit tests for the Pearson type III distribution. Stoch Environ Res Risk Assess 26:873–885. https://doi.org/10.1007/s00477-011-0519-z
Xu Y (2022) Probabilistic evaluation of the multicategory seasonal precipitation re-forecast. Meteorology 1(3):231–253. https://doi.org/10.3390/meteorology1030016

Publisher's Note