Statistical Postprocessing of Week-1 and Week-2 Precipitation Forecasts over Taiwan

HUI-LING CHANG,[a] ZOLTAN TOTH,[b] SHIH-CHUN CHOU,[a] CHIH-YUNG FENG,[c] HAN-FANG LIN,[c] PAY-LIAM LIN,[d] JING-SHAN HONG,[a] CHIN-TZU FONG,[a] AND CHIA-PING CHENG[a]

[a] *Central Weather Administration, Taipei, Taiwan*
[b] *Global Systems Laboratory, National Oceanic and Atmospheric Administration, Boulder, Colorado*
[c] *Manysplended Infotech Ltd, Taipei City, Taiwan*
[d] *Institute of Atmospheric Physics, National Central University, Jhongli City, Taiwan*

(Manuscript received 9 October 2023, in final form 14 June 2024, accepted 23 July 2024)

ABSTRACT: The predictability of precipitation is hindered by finer-scale processes not captured explicitly in global numerical models, such as convective interactions, cloud microphysics, and boundary layer dynamics. However, there is growing demand across various sectors for medium- (3–10-day) and extended-range (10–30-day) quantitative precipitation forecasts (QPFs) and probabilistic QPFs (PQPFs). This study uses a novel statistical postprocessing technique, APPM, that combines analog postprocessing (AP) with probability matching (PM) to produce week-1 and week-2 accumulated precipitation forecasts over Taiwan. AP searches for historical predictions that closely resemble the current forecast and create an AP ensemble using the observed high-resolution precipitation patterns corresponding to these forecast analogs. Frequency counting and PM are then separately applied to the AP ensemble to produce calibrated and downscaled PQPFs and bias-reduced QPFs, respectively. Evaluation over a 22-yr (1999–2020) period shows that raw ensemble forecasts from the GEFS of NOAA/NWS/Environmental Modeling Center, collected for the subseasonal experiment, are underdispersive with a wet bias. In contrast, the AP ensemble spread well represents forecast uncertainty, leading to substantially more reliable and skillful probabilistic forecasts. Furthermore, the AP-based PQPF demonstrates superior discrimination ability and yields notably greater economic benefits for a wider range of users, with the maximum economic value increasing by 30%–50% for the week-2 forecast. Compared to the raw ensemble mean forecast, the calibrated QPF exhibits lower mean absolute error and explains 3–8 times more variance in observations. Overall, the APPM technique significantly improves week-1 and week-2 QPFs and PQPFs over Taiwan.

SIGNIFICANCE STATEMENT: There are two significant challenges in improving precipitation forecasts beyond a few days in Taiwan. First, large-scale numerical models often struggle with accurately predicting precipitation locations, magnitudes, and providing sufficient detail. Second, probabilistic precipitation forecasts have been unreliable, failing to convey accurate uncertainty information to users. In response to these challenges, this study has developed a relatively simple yet effective technique that corrects the spatiotemporal distribution of predicted precipitation and downscales the forecasts from a 1° to 1-km spatial resolution. Our results demonstrate that this technique significantly alleviates these two issues, resulting in more accurate precipitation forecasts and more reliable probabilistic precipitation forecasts within a 2-week timeframe.

KEYWORDS: Precipitation; Statistical techniques; Ensembles; Probabilistic Quantitative Precipitation Forecasting (PQPF)

1. Introduction

At present, most weather forecasts are based on output from various numerical prediction models. These models predict the future atmospheric state by solving a set of governing equations, which explain changes in the atmosphere, through dynamical relationships and statistical parameterizations of finer-scale physical processes. There are several limitations that users encounter when using output from these models. First, systematic errors, which are especially common near the surface in mountainous areas,[1] grow with increasing lead time. Second, numerical models can describe the behavior of only a limited set of variables within relatively large grid boxes.

Statistical postprocessing, which primarily refers to bias correction and downscaling, can be used to address some limitations of numerical prediction models. Bias correction involves reducing or eliminating systematic errors (Mass 2003), and downscaling involves interpreting large-scale model variables in terms of small-scale user variables that are of interest to users but not directly provided by numerical prediction models in sufficient detail. These user variables include

Denotes content that is immediately available upon publication as open access.

Corresponding author: Hui-Ling Chang, lingo@cwa.gov.tw

[1] Near-surface environmental conditions (e.g., terrain and vegetation) and physical processes are usually simplified so that the surface layer of the model cannot well reflect the influence of real environment. Thus, it is difficult to correctly predict the near-surface weather conditions.

DOI: 10.1175/JHM-D-23-0177.1

pointwise daily maximum and minimum temperatures near the surface, wind gust speed, cloud cover percentage, visibility, the number of consecutive days without precipitation (Mendoza et al. 2015), and many other variables, as well.

There have been several comprehensive reviews of statistical postprocessing methods (Li et al. 2017; Vannitsem et al. 2019, 2021). Vannitsem et al. (2019) specifically concentrated on statistical postprocessing of ensemble forecasts, extending the use of statistical corrections to traditional deterministic forecasts. This book introduced important applications, particularly in hydrology, renewable energy, and long-range forecasts spanning months to decades. Li et al. (2017) conducted a thorough review of statistical postprocessing methods for meteorological and hydrological forecasts, emphasizing approaches addressing spatiotemporal and intervariable dependencies. Vannitsem et al. (2021) provided a comprehensive overview of statistical postprocessing for weather forecasts, covering theoretical advancements, challenges in operational implementation, and future prospects.

Various statistical postprocessing methods have been applied in precipitation forecasts, such as linear regression (Lu et al. 2007; Yuan et al. 2008; Chang et al. 2012), logistic regression (Wilks 2006; Wilks and Hamill 2007), artificial neural networks (Yuan et al. 2007), quantile mapping (QM; Maraun 2013; Zhao et al. 2017), a frequency matching method (FMM; Zhu and Luo 2015), and analog methods (Hamill and Juras 2006; Hamill and Whitaker 2006; Hamill et al. 2015; Horton et al. 2017, 2018). Quantile mapping (Maraun 2013) effectively corrects biases between regional climate model simulations and observations of similar resolution. However, challenges arise when applying QM to observations of much higher resolution, particularly for daily precipitation. The downscaling process introduces problems such as distorting temporal and spatial structures and overestimating area-mean extremes. This is attributed to QM's inability to introduce small-scale variability. In addition, QM is a popular method for postprocessing ensemble general circulation model forecasts (Hopson and Webster 2010). It effectively corrects bias in raw forecasts but falls short in ensuring reliability and coherence. This is because it neglects the correlation between raw ensemble forecasts and observations, resulting in negatively skillful forecasts when significant positive correlation is lacking (Zhao et al. 2017).

Analog techniques select historical analogs on the basis of a similarity criterion, either matching initial conditions (Toth 1989; Van den Dool 1989; Zorita and von Storch 1999) or forecasts from past events (Hamill and Juras 2006; Hamill and Whitaker 2006; Hamill et al. 2013, 2015). By using higher density observations, analog procedures can implicitly downscale low-resolution forecasts by imparting observed spatiotemporal variability. Specifically, analog postprocessing (AP) methods can significantly improve probabilistic precipitation forecasts (Hamill and Whitaker 2006; Hamill et al. 2015; Ben Daoud et al. 2016). Hamill and Whitaker (2006) demonstrated that incorporating both forecast precipitation and precipitable water as predictors rather than using precipitation alone can improve the precipitation forecast. In addition, long-range forecasts should have larger search areas for pattern matching than short-range forecasts. This is probably because only larger-scale features are predictable at longer forecast lead time. Additionally, matching the rank rather than the value of the mean forecast with the rank of potential analogs and spatial smoothing of resulting probability fields to reduce sampling noise slightly improves forecasting ability.

Horton et al. (2017) incorporated a moving time window (MTW) approach into AP methods, allowing a time shift between target and candidate situations. This innovation improved atmospheric circulation analogy, proving especially advantageous for heavy precipitation days. In a subsequent study, Horton et al. (2018) introduced a global optimization approach using genetic algorithms to enhance AP for precipitation forecasts. Unlike the traditional sequential method, which faces limitations in selecting optimal parameters and handling parameter dependencies, the global optimization approach automatically and objectively optimizes parameters. It considers parameter interdependencies and enables exploration of new degrees of freedom, resulting in improved forecast performance, particularly in high precipitation scenarios.

A limitation of AP methods is that sufficiently large samples of reforecasts (or hindcasts) with matching verifying observations are required to ensure analogs of good quality. However, the observed climate and forecast quality may not be stationary over long periods. Additionally, the rarer the event is, the more difficult it is to find close forecast analogs (Hamill and Whitaker 2006). Logistic regression is an alternative technique that is less prone to sampling error and generally has skill comparable to AP methods (Hamill and Whitaker 2006). Additionally, logistic regression can effectively calibrate medium-range precipitation forecasts (Wilks and Hamill 2007). However, logistic regression is more computationally expensive than AP methods because regression coefficients need to be computed individually for each precipitation threshold (Hamill and Whitaker 2006).

As precipitation often occurs on finer scales, its predictability beyond a few days is limited. However, demand for medium-range (3–10-day) and extended-range (10–30-day) precipitation forecasts (MEPFs) has increased in agriculture, forestry, livestock, and water resource management. Given the large uncertainty in precipitation forecasts, both quantitative precipitation forecasts (QPFs) and probabilistic QPFs (PQPFs) are required. QPFs are often used in sophisticated ex-ante decision-making processes. For example, a single (deterministic) or ensemble mean QPF was used as input to a hydrological model in reservoir inflow forecasting (Yang et al. 2023). This study developed a statistical postprocessing technique combining AP and probability matching (PM; Ebert 2001) to generate a more accurate single (deterministic) QPF and more reliable PQPF for medium and extended ranges.

In our study, we adopted the AP method developed by Hamill and Whitaker (2006). Some notable changes made include an extension of the forecast lead time from 6 to 14 days, the use of 7-day instead of 24-h accumulation periods, and the use of a high-resolution (1-km) observational precipitation analysis based on rain gauge data in place of a coarse-scale model reanalysis. Additionally, we considered both the mean and spread of ensemble forecasts when selecting analogs and generated postprocessed QPF in addition to PQPF. Our

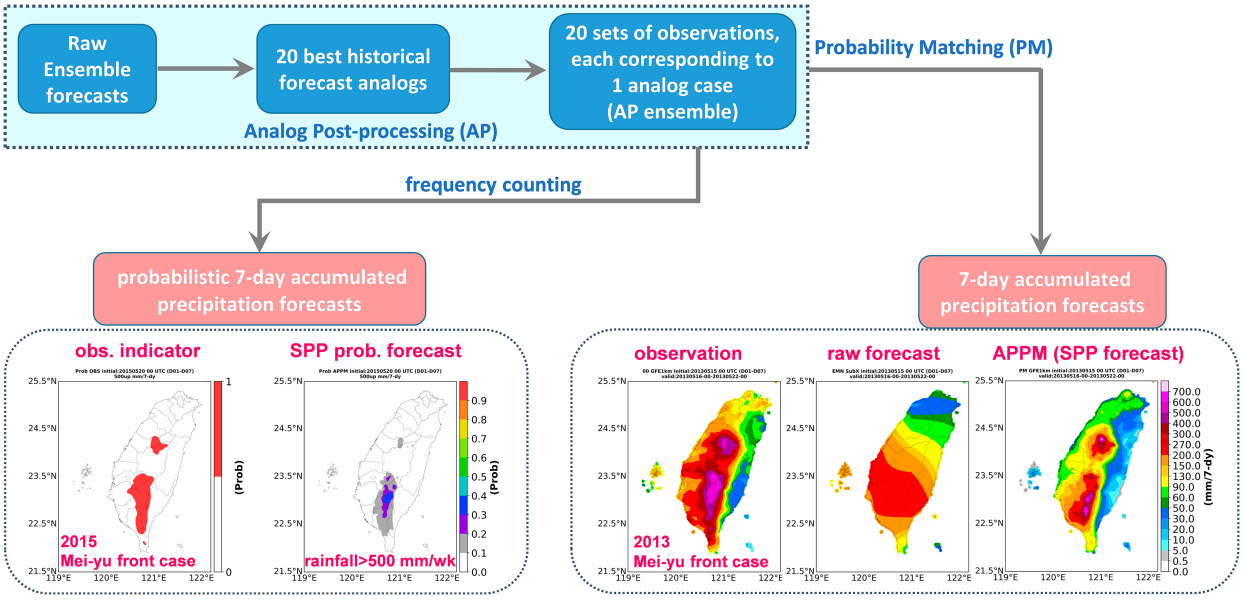

FIG. 1. Statistical postprocessing procedure of probabilistic and single (deterministic) precipitation forecasts.

approach is to predict precipitation conditioned on the forecast of the large-scale circulation which still retains predictability beyond a few days (i.e., conditional climatic distribution), instead of a single prediction of individual precipitation events.

This paper is organized as follows. The data sources and validation methodology are introduced in section 2. The statistical postprocessing methodology is described in section 3. Forecast evaluation results and discussion are presented in sections 4 and 5, respectively. Finally, the study is concluded in section 6.

2. Data sources and validation

This study uses both retrospective (i.e., reforecast) and operational precipitation forecasts for lead times up to 14 days from the National Oceanic and Atmospheric Administration/National Weather Service/Environmental Modeling Center (NOAA/NWS/EMC) Global Ensemble Forecast System (GEFS) collected for the subseasonal experiment (SubX; Pegion et al. 2019), over the period of January 1999–September 2020. EMC-GEFS is an operational model; thus, the real-time operational forecasts and reforecasts were generated using the same model. The number of ensemble members for the reforecast (January 1999–December 2016) and real-time operational periods (August 2017–September 2020) is 10 and 20, respectively. Over both periods, the update frequency is once per week, with forecasts initialized at 0000 UTC every wednesday. The horizontal resolution is $1^\circ \times 1^\circ$ in latitude and longitude.

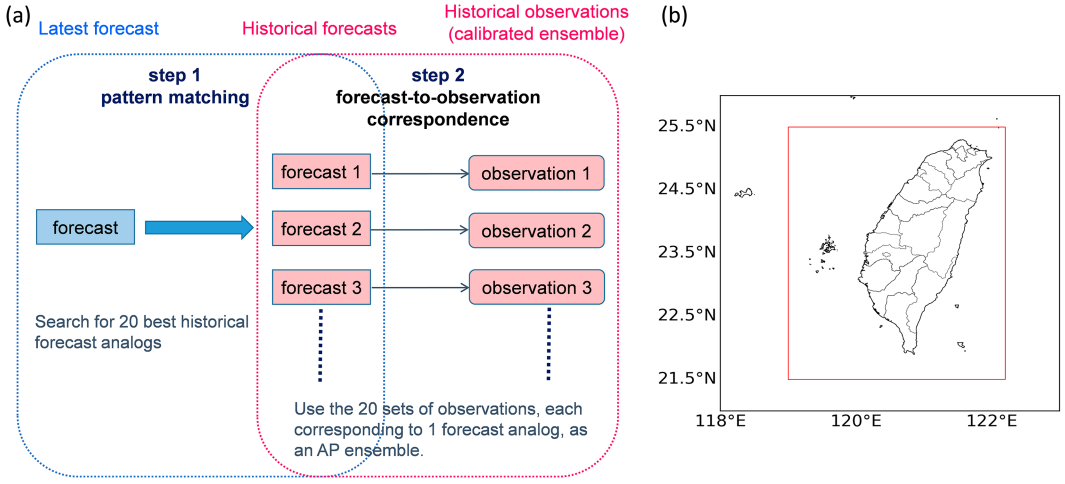

FIG. 2. (a) Schematic diagram of AP and (b) search area for pattern matching (black outer box) and verification area (Taiwan Island and its outer islands inside the red box).

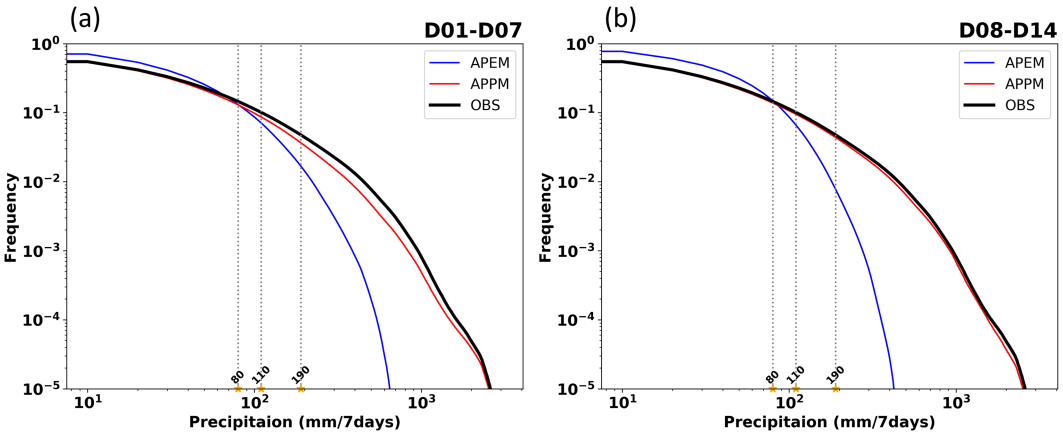

FIG. 3. Frequency distributions of 7-day accumulated precipitation evaluated over the entire 22-yr sample for (a) 1–7- and (b) 8–14-day lead times for APEM (blue curve), APPM (red curve), and observation (black curve). Vertical dotted lines from left to right mark accumulated precipitation at the 85th, 90th, and 95th percentiles.

For training and validation, a gridded precipitation analysis based on rain gauge data for the entire experimental period was prepared using a variant of the simple kriging method (Ali et al. 2005). The horizontal resolution of the observational precipitation analysis is 1 km. Hence, beyond bias correction, AP also downscales the forecasts from 1° to 1-km spatial resolution. As the ultimate source of finer-scale information, the observational precipitation analysis plays a critical role in the calibration of precipitation forecasts. Detailed information on gauge spacing and estimation error is provided in appendix A.

To increase the number of samples, validation was performed using a leave-one-out cross-validation procedure (Wilks 2011).

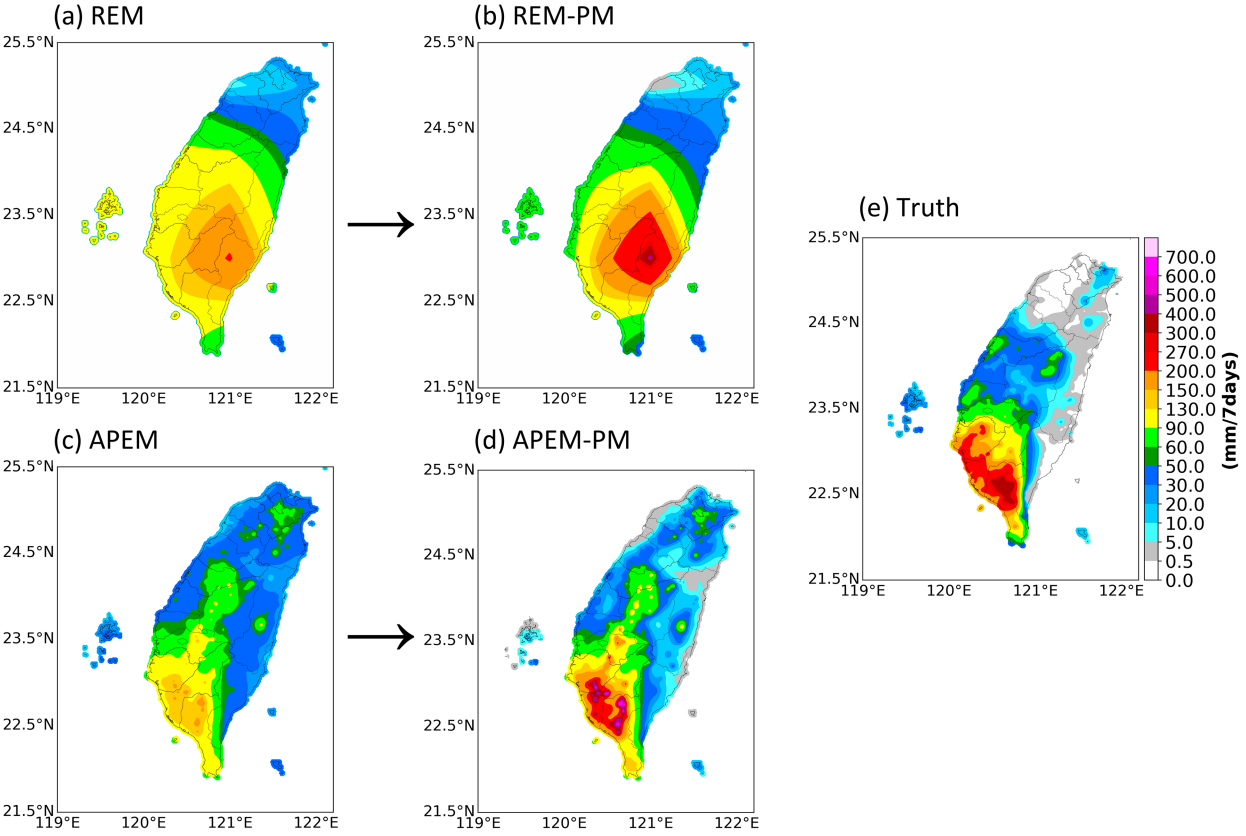

FIG. 4. 1–7-day QPF [(a) RAW^3 ensemble mean (REM), (b) REM-PM, (c) APEM, and (d) APEM-PM] and (e) observationally based analysis for the period ending at 0000 UTC 11 Jul 2019 (Typhoon Mun).

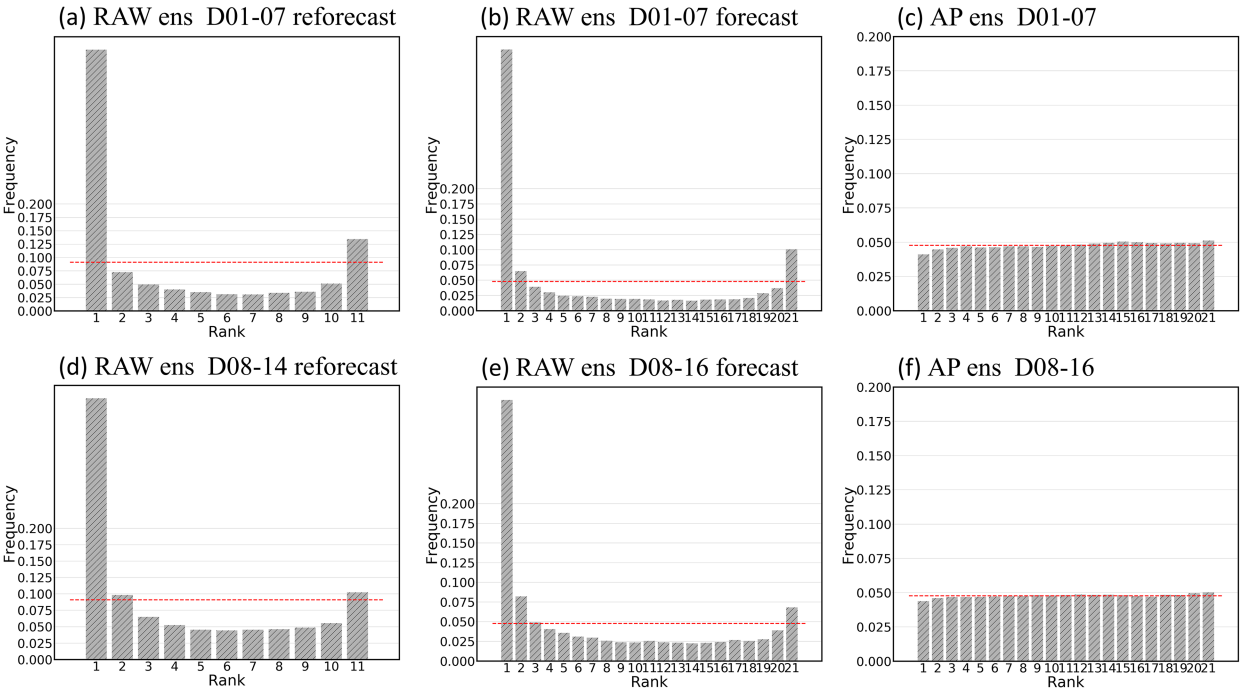

FIG. 5. (top) 1–7- and (bottom) 8–14-day analysis rank histograms for (a),(d) RAW ensembles for reforecast, (b),(e) RAW ensembles for real-time forecast, and (c),(f) AP ensembles. Horizontal dashed lines denote frequency for uniform rank distribution.

AP was separately trained to calibrate forecasts made in each year of the full sample. As an example, for calibrating reforecasts issued in 2005, training was based on data from 1999 to 2004 and 2006 to 2022. In addition, Taiwan has three distinct weather regimes: summer (July–September), the winter half year (October–April), and the mei-yu[2] season (15 May–15 June). To address seasonal variations, training was conducted separately for three slightly overlapping periods: summer (16 June–15 October), the winter half year (15 September–15 May), and the mei-yu (15 April–15 July) season.

3. Methodology

This study produced calibrated 7-day accumulated precipitation forecasts, including probabilistic and single (deterministic) forecasting, using statistical postprocessing methods. The statistical postprocessing procedures are shown in Fig. 1. First, AP (Hamill and Whitaker 2006; Hamill et al. 2015) was used to obtain historical forecast analogs that most resemble the current ensemble forecast. We searched for the 20 past dates where the ensemble forecast was most similar to the ensemble forecast of today. The observations corresponding to the best forecast analogs then served as a calibrated forecast ensemble (referred to as the AP ensemble hereafter). Subsequently, PM was applied to the AP ensemble to derive a single (deterministic) forecast (called the APPM forecast), and frequency counting was used to obtain probabilistic forecasts. That is, the PQPF was generated using the relative frequency of the event in the ensemble. For example, if 5 of the 20 ensemble members at a grid point indicated greater than 50 mm of precipitation, the probability of that event was 25% (5/20).

a. AP

Pattern matching between the current forecast and historical forecasts was conducted to identify the 20 closest historical forecast analogs. The 20 sets of observational precipitation analysis corresponding to the forecast analogs were used as a calibrated forecast ensemble (AP forecast ensemble), representing possible observed weather scenarios (Fig. 2a). The distance-based similarity criterion $D(t)$ [Eq. (1)], including ensemble mean $\bar{x}$ [Eq. (2)] and ensemble spread [Eq. (3)] terms, is defined as follows:

$$D(t) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(\bar{x}^{l,tc} - \bar{x}^{l,t})^2 \frac{1}{L}\sum_{l=1}^{L}(s^{l,tc} - s^{l,t})^2}, \quad (1)$$

$$\bar{x} = \frac{1}{M}\sum_{m=1}^{M} x_m, \quad (2)$$

$$s = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(x_m - \bar{x})^2}, \quad (3)$$

[2] Mei-yu season is the East Asian rainy season, also known as the plum rain. The term "mei-yu" is derived from the Chinese pronunciation, signifying the rain that falls during the ripening period of plums. It is caused by precipitation along a persistent stationary front, referred to as the mei-yu front, over Taiwan from mid-May to mid-June. The wet season ends during the early summer when the subtropical ridge strengthens enough to push this front north of the region. These weather systems can result in heavy precipitation and flooding.

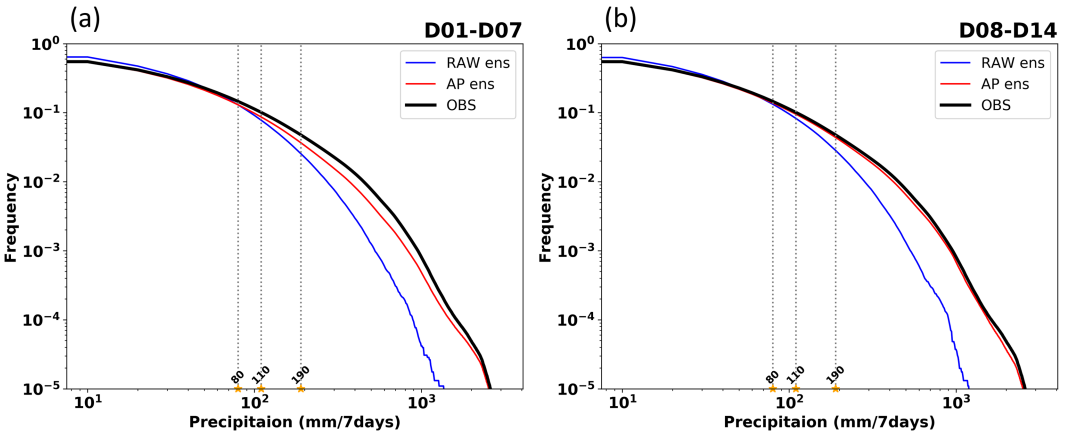

FIG. 6. As in Fig. 3, but for RAW ensemble (blue curve), AP ensemble (red curve), and observation (black curve).

where x is the forecast variable, which is 7-day accumulated precipitation in this study. Let tc be the current date and t be the chosen date in the historical dataset whose forecast data will be compared against the forecast at tc. The L is the number of grid points in the search area for pattern matching (Fig. 2b), and M is the number of ensemble members. The search domain is much larger than the verification domain, which covers only the Taiwan land area. This is because large-scale information is essential for pattern matching, given the consideration of predictability.

As the range of predictable scales shrinks with increasing lead time (Boer 2003), forecast skill also strongly depends on spatial scales. In the extended range, only slowly changing larger-scale features are predictable. The 7-day accumulated precipitation, which is both a predictor and predictand in this study, reflects slowly changing large-scale conditions and

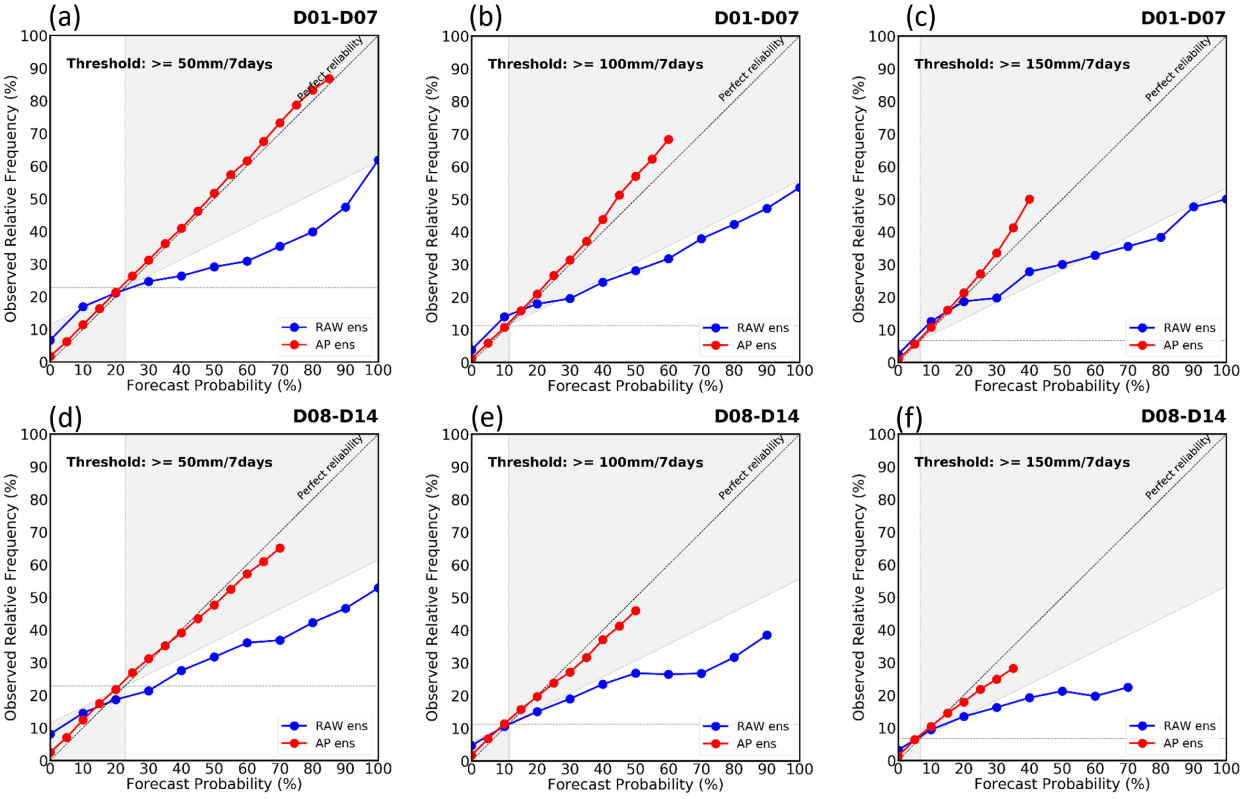

FIG. 7. Reliability diagrams for RAW (blue) and AP (red) PQPFs for (a),(d) 50, (b),(e) 100, and (c),(f) 150 mm week^{-1} for (top) 1–7-day and (bottom) 8–14-day lead times, evaluated over the entire 22-yr sample. Horizontal dashed line indicates sample climatological frequency.

consequently has some predictability in the extended range. The basic results from our study remain unchanged when 7-day accumulated precipitation as a predictor is replaced with large-scale circulation indices (see appendix B) in the search for analogs (not shown).

b. PM

PM (Ebert 2001; Su et al. 2014) was used to convert the AP ensemble forecasts into a single (deterministic) forecast after the spatial distribution of precipitation forecasts was corrected by AP. For each forecast, consider the cumulative distribution functions of two components: 1) the ensemble mean (EM) values at each grid point within a specified domain and 2) all perturbed members at each grid point across the same domain. The PM value for each EM value (i.e., probability matched mean) is determined by identifying the precipitation value in distribution 2 that corresponds to the cumulative frequency value of the EM precipitation value in distribution 1. During the probability matching process, only the current PM ensemble forecasts were used and historical forecast and corresponding observational precipitation analysis were not required. Therefore, the PM is not a downscaling method.

In essence, when applied to the mean of the AP ensemble, PM relabels precipitation forecast values by using aggregated ensemble member data to better match the frequency distribution of forecast precipitation with corresponding observations. The effect is illustrated in Figs. 3a and 3b. The AP ensemble mean (APEM; blue curve) has lower and higher frequency for heavy and light precipitation than the observation, respectively, that is, because the averaging process in computing an ensemble mean filters out extreme values and coarsens the data. Such a phenomenon is mitigated in APPM (red curve)—its frequency distribution of accumulated precipitation was very close to that of the observations. In other words, the range of forecasted values was close to that in the observational dataset.

TABLE 1. 80th, 85th, 90th, and 95th climatological percentile of observed 7-day accumulated precipitation over Taiwan for winter half year, mei-yu season, and summer. The percentile precipitation values are rounded to the nearest 5 mm week^{-1}. The total number of 7-day periods between January 1999 and September 2020 contributing to the statistics is indicated in parentheses.

Accumulated precipitation (mm week^{-1})	Winter half year (636 cases)	Mei-yu (184 cases)	Summer (283 cases)
80th percentile	30	100	100
85th percentile	40	125	150
90th percentile	50	150	210
95th percentile	95	240	330

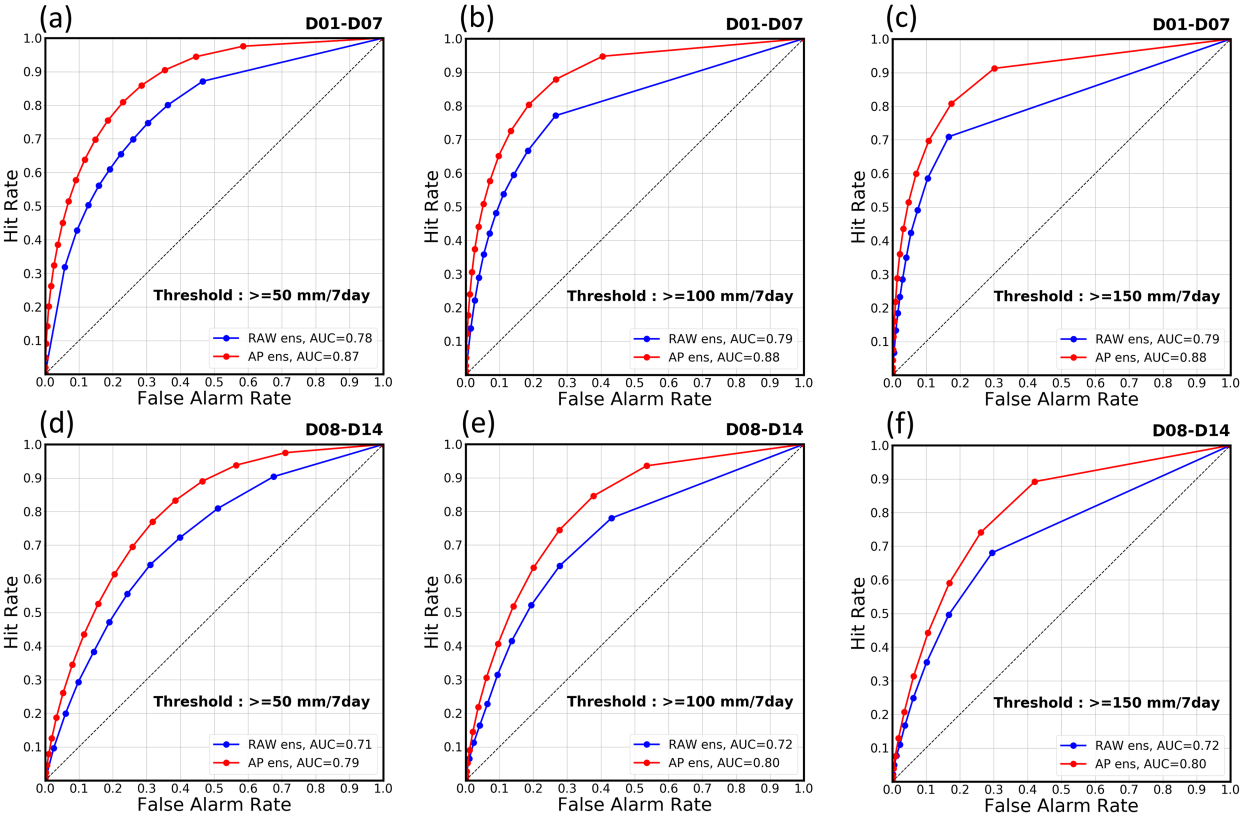

FIG. 8. As in Fig. 7, but for ROC curves.

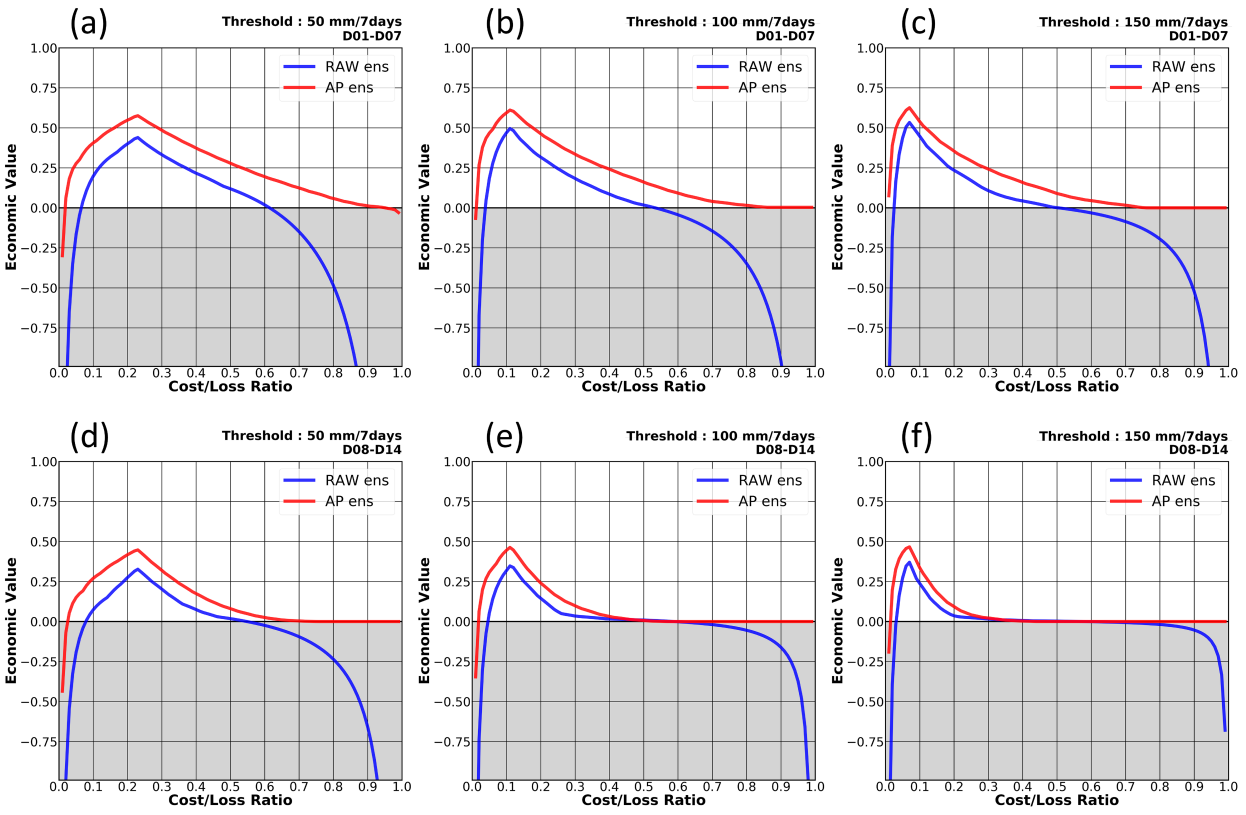

FIG. 9. As in Fig. 7, but for potential EV curve as a function of a user's cost/loss ratio.

As demonstrated by the case of Typhoon Mun in 2019 (Fig. 4), PM only corrects the frequency distribution of precipitation over the entire domain; however, it fails to address the issue of misplaced precipitation caused by the model's coarse representation of orography and other limitations (cf. Fig. 4a with Fig. 4b). On the other hand, AP corrects the positional error (Fig. 4c) and, with a PM application (Fig. 4d), produces a forecast with a realistic range of precipitation amounts (cf. Fig. 4d with the corresponding observationally based analysis in Fig. 4e).

4. Forecast evaluation results

a. *Distribution of ensembles*

Analysis rank histograms (Hamill 2001), which are also called Talagrand diagrams, can be used to evaluate whether the spread of ensemble forecasts adequately represents the underlying uncertainty. The results of an assessment of analysis rank histograms indicated that the RAW ensemble forecast was underdispersive (Figs. 5a,b,d,e). Approximately 33%–49% of the samples had observed precipitation values that were lower than the entire range of ensemble forecasted values (falling into rank 1), whereas 7%–14% of the samples had observed values that were higher than the forecasted values of all members (falling into the rightmost 11th or 21st ranks). In contrast, the distribution of the downscaled AP ensemble was calibrated with most of the bias removed (Figs. 5c,f).

The frequency distribution of accumulated precipitation for the RAW ensemble indicated overforecasting for light precipitation and underforecasting for heavy precipitation (Fig. 6), consistent with an overly high number of verifying cases falling into rank 1[3] (and to some extent, into the rightmost 11th or 21st ranks of the analysis rank histograms) and overly low number of cases falling into the middle ranks (Figs. 5a,b,d,e). In other words, the RAW ensemble produced a narrower range of precipitation values than the observed climatology; therefore, it cannot contain the right tail of the precipitation spectrum. By contrast, frequencies for the AP ensemble were much closer to observed values compared with those for the RAW ensembles (Fig. 6). In other words, the calibrated AP ensemble better represents the precipitation variability of the observed climatology.

b. *7-day probabilistic quantitative precipitation forecasts*

Statistical reliability assesses systematic error (biases), whereas statistical resolution measures predictive skill in probabilistic forecast systems (Toth et al. 2006). Reliability and resolution are independent attributes of forecast systems. For probabilistic forecasts, reliability assesses how well the

[3] For rank 1, a few samples (less than 0.5%) have zero precipitation when all ensemble members forecast nonzero precipitation. Such a situation is common in other models. However, this phenomenon is not the main reason for the overly high frequency of rank 1.

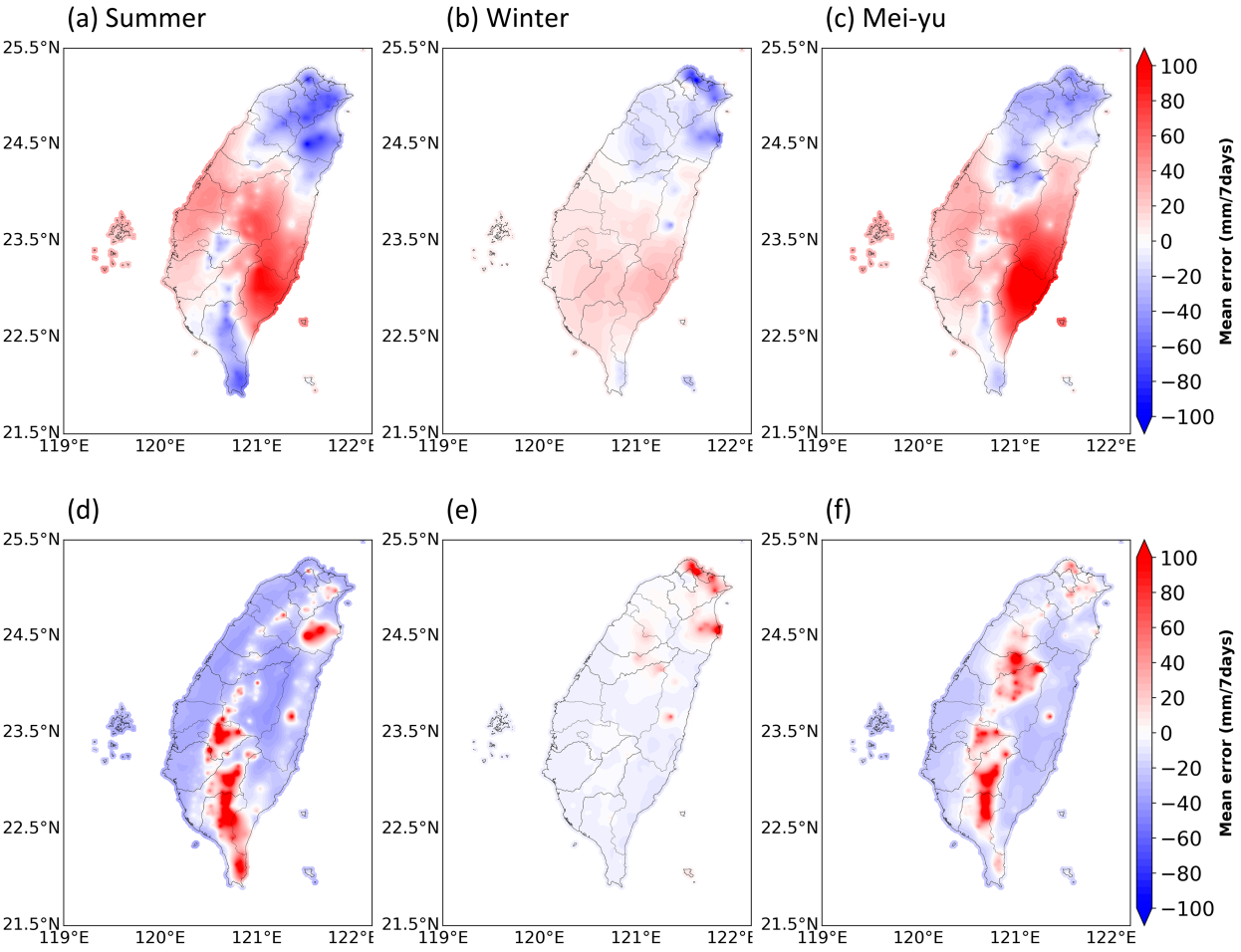

FIG. 10. Spatial distribution of ME from the 1–7-day QPFs for (top) REM and (bottom) APPM for the (a),(d) summer, (b),(e) winter half year, and (c),(f) mei-yu seasons, evaluated over the entire 22-yr sample.

forecast probabilities of an event correspond to the conditionally observed frequencies of the predicted event. The proximity of the observed frequency line to the diagonal line on reliability diagrams provides a graphical assessment of reliability (Hsu and Murphy, 1986). The reliability diagrams of the RAW and AP PQPFs at different thresholds and lead times are shown in Fig. 7. The RAW PQPF had no forecast skill (Brier skill score < 0; as indicated by the reliability curve not falling in the gray area) at almost every tested forecast probability threshold, whereas the AP PQPF had forecast skill at every probability level. The RAW probability forecast values were typically higher/lower than observed frequency for probability values above/below the climatological frequency of the predicted events (indicated by the horizontal line in each panel of Fig. 7). Precipitation events of 100 and 150 mm week^{-1} were rare in the winter half year, with the latter being at approximately the 95th percentile of observed climatology in the mei-yu season and the 90th percentile in the summer season (Table 1). Even for such heavy and rare precipitation events as 150 mm week^{-1}, the calibrated PQPF had better reliability compared with the RAW forecasts (Figs. 7c,f).

One measure of the predictive skill used in this study is relative operating characteristic (ROC; Mason and Graham 1999; Jolliffe and Stephenson 2003). A forecast system with good discrimination between observed events and nonevents has an evaluation curve that is close to the upper-left corner of an ROC plot. The ROC curves shown in Fig. 8 indicate that the PQPF calibrated by AP had higher skill in discrimination than the RAW PQPF. This was because precipitation patterns were corrected in the AP ensemble that in turn improved the predictive skill in discrimination.

A system with higher predictive skill may benefit some users more than others. This can be assessed by the potential economic benefit in decision-making processes of users (Murphy, 1977; Katz and Murphy, 1997; Zhu et al. 2002). Richardson (2000) defines relative economic value (EV) of a forecast system as the reduction in expected expenses over the use of climatological information relative to the reduction that would be obtained by using a perfect forecast. Potential EV is then defined as relative EV aggregated over the entire range (0–1) of cost/loss ratios assuming that to maximize

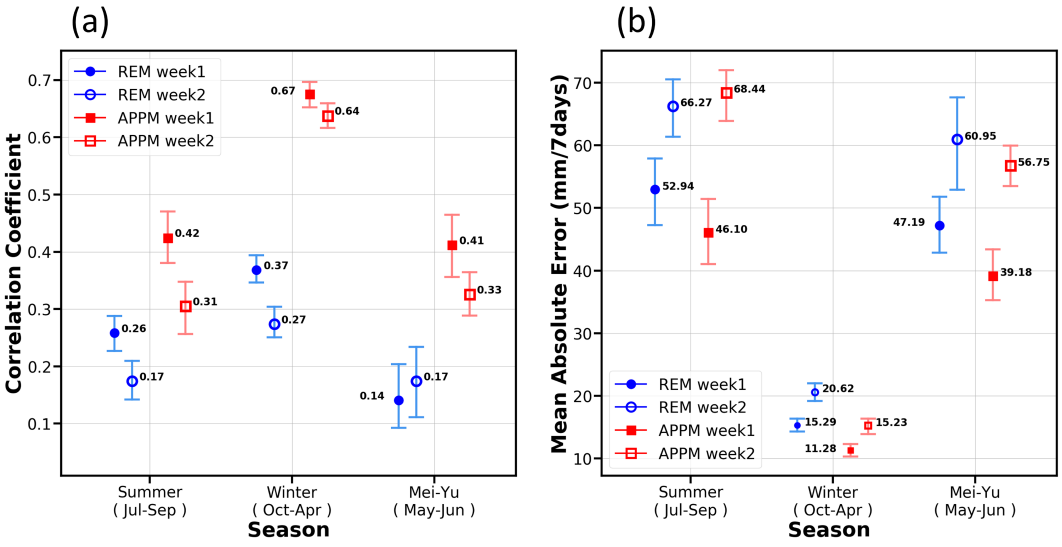

FIG. 11. (a) Correlation coefficient and (b) MAE calculated using forecast and observed precipitation. Data in (a) and (b) are presented as median values with 95% confidence intervals based on 10 000 bootstrap samples for the summer, winter half year, and mei-yu seasons, evaluated over the entire 22-yr sample. Confident intervals represented by the blue line with a circle mark and the red line with a square mark correspond to results from REM and APPM QPF, respectively. Solid and hollow marks denote confidence intervals for the 1–7-day and 8–14-day QPFs, respectively.

their relative EV, users base their decisions on the optimal probability threshold corresponding to their specific cost/loss ratio. With such a strategy, users exploit the full discrimination ability of the system they use (Chang et al. 2015). An analysis of relative EV (Fig. 9) reveals that for any threshold and at any lead time, more users (i.e., a much wider spectrum of cost/loss ratio) can obtain more benefits from the calibrated PQPF as compared to the RAW PQPF, with a generally higher gain in decision-making. Greater economic benefit is indicated by a cost/loss ratio closer to 0 or 1. The relative EV of those users turns from negative to positive if they change from using a RAW to using an AP PQPF as the basis for their decision-making.

c. *7-day quantitative precipitation forecasts*

1) ADVANCEMENT OF APPM OVER REM

When evaluated statistically, APPM outperformed RAW ensemble mean (REM). Figure 10 illustrates the spatial distribution of forecast biases for week-1 QPFs from APPM and REM in different seasons. Overall, while APPM exhibits obvious overforecasting in regions experiencing heavy rainfall across various seasons, its mean error (ME) is much closer to zero compared to REM. In other words, APPM improves the forecast bias compared to raw forecasts. The precipitation areas in each season primarily reflect terrain effects, with heavy precipitation occurring on the windward side of the mountains. In summer and winter, Taiwan is influenced by the southwest monsoon and the northeast monsoon, leading to heavy precipitation in the mountainous regions of southwestern Taiwan and the northeastern corner, respectively. In contrast, due to the insufficient resolution of global models leading to the central mountain range being smoothed and inaccurately positioned, the spatial distribution of mean error from the REM QPF does not realistically reflect the terrain effects.

Considering the uncertainty associated with sampling variability and the limitations in sample size, 95% confidence intervals for the median of the correlation coefficient and mean absolute error (MAE) for the entire evaluation period were constructed through bootstrap resampling with replacement (Chu 2002; Wilks 2011; Chang et al. 2017). The APPM QPF had significantly better precipitation patterns than did the REM QPF (Fig. 11a). For all seasons and lead times (except the MAE for week-2 forecast in summer), the 95% intervals for APPM and REM did not overlap, indicating that the AP algorithm significantly improved forecast performance (higher correlation and lower MAE; Fig. 11). This was further highlighted by the 3- to 8-fold increase in the variance explained by the APPM forecasts[4] in the observed precipitation, as compared with the REM forecasts (values in bold, Table 2).

2) IMPROVEMENT OF APPM OVER APEM FOR HEAVY PRECIPITATION EVENTS

As cross-sectoral users are more concerned with heavy or extreme precipitation events compared to light or moderate precipitation events, bias ratios (Wilks 2011) at the 80th, 85th,

[4] A boxplot analysis (not shown) revealed that the distributions of correlations coefficients were asymmetrical; therefore, the median rather than the mean was used as the best estimate to calculate the observation variance explained by the forecasts (Chang et al. 2017).

TABLE 2. Explained variance[5] for different seasons and lead times, along with the ratio of APPM to REM values (bold), evaluated over the entire 22-yr sample.

Percentage of explained variance		Summer (285 cases)	Winter (636 cases)	Mei-yu (183 cases)
1–7 day	REM	6.8%	13.7%	2.0%
	APPM	17.6%	44.9%	16.8%
	APPM/REM	**2.6**	**3.3**	**8.4**
8–14 day	REM	2.9%	7.3%	2.9%
	APPM	9.61%	41.0%	10.9%
	APPM/REM	**3.3**	**5.6**	**3.8**

and 90th climatological percentiles (Fig. 12) are shown to further explain why we opted for the PM instead of the ensemble mean as a single-value forecast. In the summer and mei-yu seasons, during which Taiwan frequently encounters substantial precipitation, the APEM QPF exhibits underforecasting beyond the 80th percentile. Moreover, the dry bias becomes increasingly pronounced as climatological percentiles rise. The bias ratios of APEM, approaching zero at the 85th and 90th percentiles, show that APEM rarely issues heavy or extreme precipitation forecasts. In contrast, the APPM QPF exhibits bias ratios closer to 1 (indicating unbiasedness) at each threshold compared to the APEM QPF.

3) COMPARISON WITH THE PREVIOUS OPERATIONAL CALIBRATION METHOD: FMM

The Central Weather Administration (CWA) in Taiwan has been actively promoting climate services in recent years, tailoring a variety of customized forecast products to meet the needs of users across different sectors. In the fields of agriculture and water resource management, there has been a rapid increase in demand for MEPFs. Before the APPM was implemented operationally in 2021, the CWA utilized the FMM (Zhu and Luo 2015) to calibrate model bias in MEPFs.

The FMM aims to eliminate the frequency bias in forecasts by aligning the frequency distribution of each forecast with that of the verifying analysis (Fig. 13). Both the forecast and corresponding observed cumulative distribution functions (CDFs) are updated using the decaying average method (Cui et al. 2012) as follows:

$$\overline{\mathrm{CDF}_i}(t) = (1 - w) \times \overline{\mathrm{CDF}_i}(t - 1) + w \times \mathrm{CDF}_i(t), \qquad (4)$$

where $\overline{\mathrm{CDF}_i}(t)$ represents the decaying averaged CDF at threshold i for the current date t, while $\overline{\mathrm{CDF}_i}(t-1)$ is the prior decaying averaged CDF at threshold i for the previous date $t - 1$. In this study, $t - 1$ refers to the date 7 days ago. The most recent CDF at threshold i for the current date t is denoted as $\overline{\mathrm{CDF}_i}(t)$, and w is the decaying weight within the range of 0 and 1. This weight controls how much influence or contribution we give to $\overline{\mathrm{CDF}_i}(t)$ from both $\overline{\mathrm{CDF}_i}(t-1)$ and $\overline{\mathrm{CDF}_i}(t)$ Sensitivity tests are conducted to determine the optimal decaying weight. For instance, the optimal weight for winter half-year forecasts has been found to be 0.05. This implies that approximately 20 weeks of historical data ($1/0.05 = 20$) from the target forecast date are used to calculate a weighted average of CDF values over the land area of Taiwan.

This northeast monsoon precipitation case was also used to illustrate the limitation of the FMM in bias correction. The precipitation pattern in the REM forecast was noticeably different from the actual precipitation pattern (Fig. 14b vs Fig. 14e). This discrepancy can be attributed to the presence of the Central Mountain Range, which covers a vast portion of the land area of Taiwan (Fig. 14d). Consequently, the spatiotemporal distribution (or pattern) of precipitation across Taiwan area is primarily driven by interactions between complex orography and surrounding large-scale circulations. As orography in global models is quite smooth (Fig. 14a), raw forecasts often exhibit unrealistic precipitation patterns. Unfortunately, the FMM can only correct pointwise amounts of precipitation in raw forecasts, while ignores changes in its spatiotemporal distributions (Fig. 14c). In other words, although calibration methods such as the FMM adjust the precipitation amount at each grid point, they do not improve overall spatial patterns of precipitation. This limitation severely affects statistical postprocessing applications in Taiwan. In contrast, the APPM method not only corrects the precipitation pattern and magnitude but also provides fine-scale details of precipitation (Fig. 14f).

5. Discussion

a. *Sensitivity experiments on AP*

1) IMPACT OF SIMILARITY CRITERION

Ensemble forecasts offer case-dependent forecast uncertainty information to users, conveyed through the spread of the ensemble (Toth et al. 2001). An ideal ensemble forecast system should have a good spread–skill relationship, that is, having the same magnitude of ensemble spread as their forecast error at the same lead time to effectively represent forecast uncertainty (Kalnay and Dalcher 1987; Zhu 2005). Unfortunately, most ensemble forecasts are underdispersive; the spread, however, typically rises with increased forecast error.

In analog postprocessing, generally only the ensemble mean forecast is used for pattern matching (Hamill and Whitaker 2006; Hamill et al. 2015). To assess the impact of ensemble spread on analogs, we introduced the ensemble spread term into the similarity criterion to investigate whether better analogs could be identified through the incorporation of this forecast uncertainty constraint. Sensitivity results indicate

[5] In statistics, explained variance measures the proportion to which a model accounts for the variance of a given dataset. In Table 2, "the percentage of explained variance" means "the percentage of observed precipitation variance explained by the APPM or REM." In a linear regression model (such as the one used here to assess the linear relationship between observed and forecast precipitation), the explained variance is quantified by R-squared (R^2), which is equal to the square of the correlation coefficient.

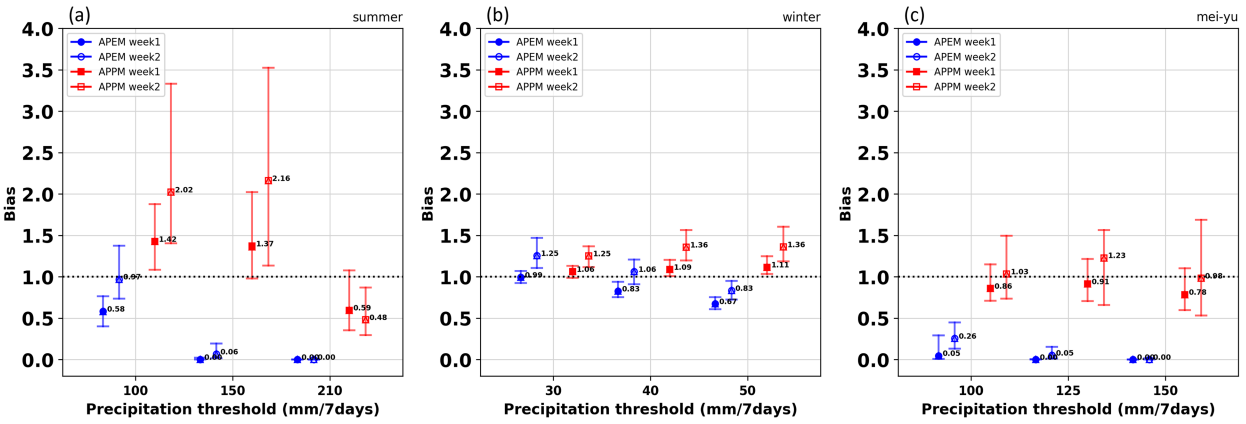

FIG. 12. As in Fig. 11, but for bias ratios evaluated for (a) summer, (b) winter half year, and (c) mei-yu seasons, for three precipitation thresholds along the x axis, from left to right: the 80th, 85th, and 90th climatological percentiles of observed 7-day accumulated precipitation over Taiwan. Confidence intervals represented by the blue line with a circle mark and the red line with a square mark correspond to results from APEM and APPM QPF, respectively.

that integrating the ensemble spread term into the distance-based similarity criterion led to an improvement in the accuracy of the accumulated precipitation frequency distribution for the AP ensemble. Additionally, this incorporation marginally increased the reliability, discrimination ability, and economic value of the AP-based PQPFs.

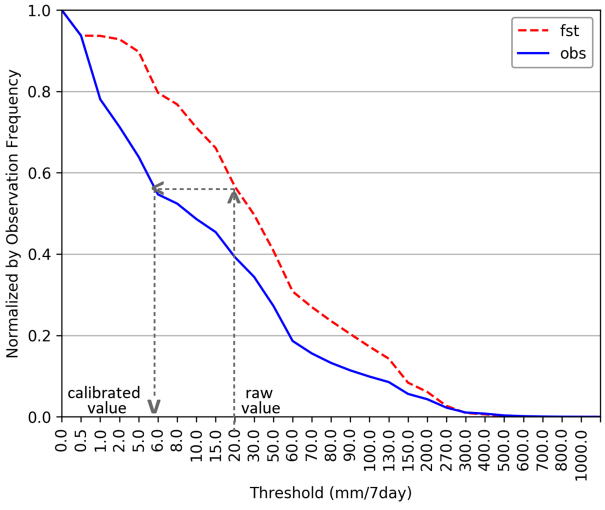

FIG. 13. Illustration of the FMM using a northeast monsoon precipitation case. The red and blue curves are the CDFs at various precipitation thresholds, denoted as $\overline{\mathrm{CDF}}_i(t)$ in Eq. (4), for a 1–7-day QPF and the corresponding observation starting at 0000 UTC 26 Jan 2000. These two curves depict the occurrence frequency of events with 7-day accumulated precipitation greater than each threshold on the x axis and were constructed using the samples from all grids over Taiwan Island and its outer islands (Fig. 2b). The forecast datasets used for forecast and observation are the same as those used for the APPM in this study. Note that the two curves overlap below the 0.5-mm threshold in this case. The forecast CDF (red curve) will be adjusted to align with the observed CDF (blue curve) through the FMM. Refer to the main text for further details.

2) IMPACT OF THE NUMBER OF ANALOGS

According to Hamill and Juras (2006), the number of analogs should vary as a function of both forecast lead time and how extreme the precipitation forecast is. The number of analogs should be smaller for extreme cases and shorter lead times, but larger for common cases and longer lead times. Sensitivity experiments were conducted using different numbers of analogs (10, 15, 20, and 25 for week 1 and 15, 20, 25, and 30 for week 2), and the results from evaluation of rank histogram, reliability diagram, ROC, and potential EV showed that adopting 20 analogs was the optimal choice for 1–7- and 8–14-day accumulated precipitation forecasts.

3) IMPACT OF REFORECAST LENGTH

One limitation of AP methods is the necessity for sufficiently large samples of reforecasts with matching verifying observations to ensure high-quality analogs (Hamill and Whitaker 2006). To assess the impact of reforecast length on AP results, we divided the entire 22-yr sample into training (before 2013) and validation periods (2013–20). In a sensitivity experiment, we varied the length of the training period among three scenarios: 1) AP-T14y with a 14-yr training period from 1999 to 2012, 2) AP-T7y with a 7-yr training period from 2006 to 2012, and 3) AP-T3y with a 3-yr training period from 2010 to 2012.

The results from the sensitivity experiment reveals that an insufficient length of reforecasts (AP-T3y) is not suitable for AP, as it adversely affects the reliability and discrimination ability of calibrated probabilistic forecasts. Comparing AP-T7y and AP-T14y indicates that doubling the training sample size from 7 to 14 years may not significantly enhance the results, as the benefits have already been achieved with a 7-yr sample size. A 7-yr reforecast sample is sufficient for achieving good-quality bias correction, resulting in performance close enough to optimal (Fig. 15).

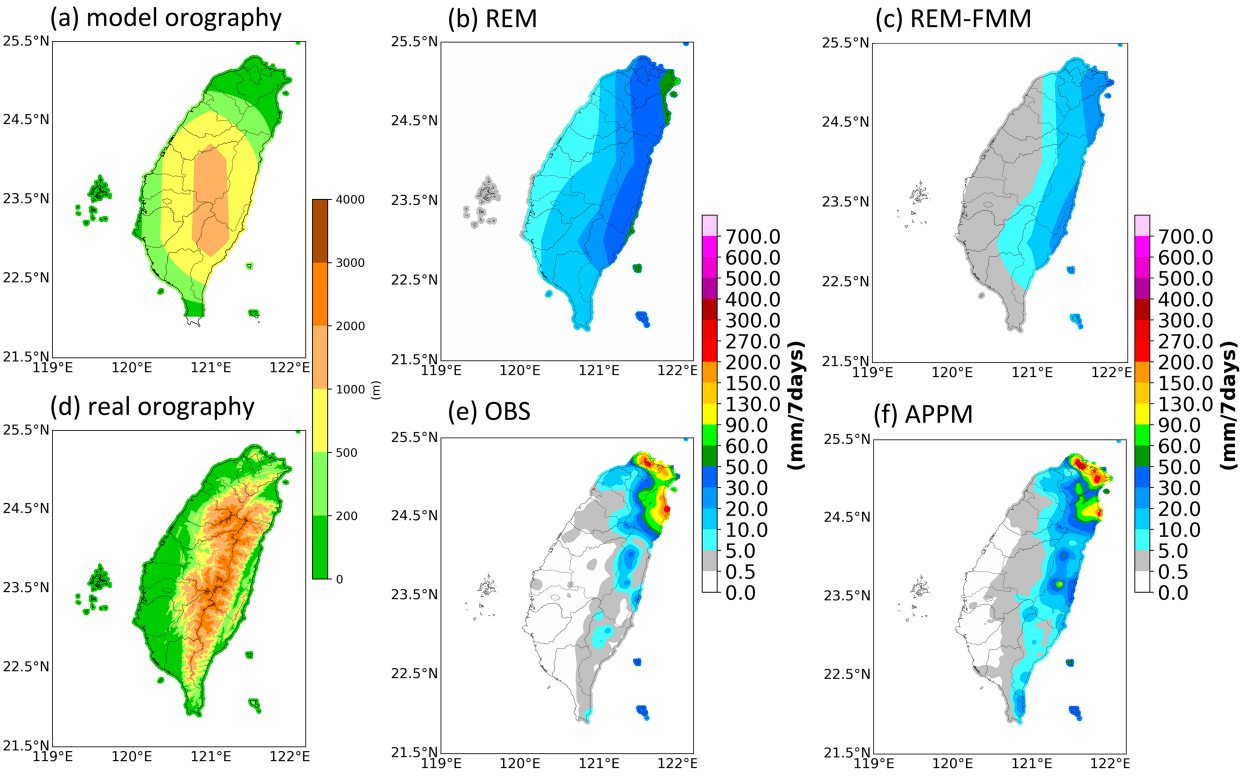

FIG. 14. (a) Model orography from the EMC -GEFS for the SubX and 1–7-day QPF ending at 0000 UTC 3 Feb 2000 (northeast monsoon) for (b) REM, (c) REM-FMM, and (f) APPM. (d) Real orography and (e) corresponding observed precipitation analysis for the same case.

b. Unique contribution of the research

While analog postprocessing (AP) and probability matching (PM) have been documented in prior research, our unique contribution lies in their application to address specific challenges related to MEPFs within the complex geographical context of Taiwan. Our study addresses three key challenges:

1) PRODUCING OBSERVATIONAL ANALYSES REFLECTING REAL PRECIPITATION SCENARIOS

To generate well-calibrated high-resolution MEPFs aligned with user demands, a comprehensive set of high-resolution precipitation analyses is crucial. We abandoned the existing observational analyses–Taiwan Station-based Analysis (TaiSA v1.0), which utilized a consistent number of rain gauge observations for homogeneity in climatology analysis. Instead, we produced a new set of precipitation analysis based on all rain gauge observations across Taiwan, ensuring a reflection of real precipitation scenarios under complex terrain (refer to appendix A for detailed information).

2) DETERMINING CRITICAL PREDICTORS FOR MEPFS

Initially, three sets of large-scale circulation indices were employed as predictors for pattern matching in the AP process (see appendix B). Through sensitivity experiments, we discovered that using 7-day accumulated precipitation as a predictor yields comparable or even slightly better forecast quality. This simplicity provides a more straightforward predictor with predictability for MEPFs.

3) INTEGRATING AP AND PM TO PRODUCE REALISTIC QPFS

Unlike prior AP studies focusing solely on probabilistic forecasts, our research combines AP and PM methods to derive both PQPFs and QPFs. This integration is vital to meet the urgent demand for realistic QPFs in water resources management, as it ensures accurate spatiotemporal distributions of precipitation and a realistic range of precipitation amounts.

6. Conclusions

In Taiwan, MEPFs lack accuracy in predicting precipitation locations and magnitude and provide insufficient detail. This is mainly due to the fact that strong orographic effects are not well captured by large-scale numerical models. In addition, the underdispersion of ensemble members limits their ability to adequately represent forecast uncertainty. In this study, we integrate the AP and PM methods to mitigate these problems and produce calibrated and downscaled QPFs and PQPFs.

We found that the calibrated forecast corrected the magnitude and large-scale distribution of precipitation and also provided valuable small-scale details influenced by orography.

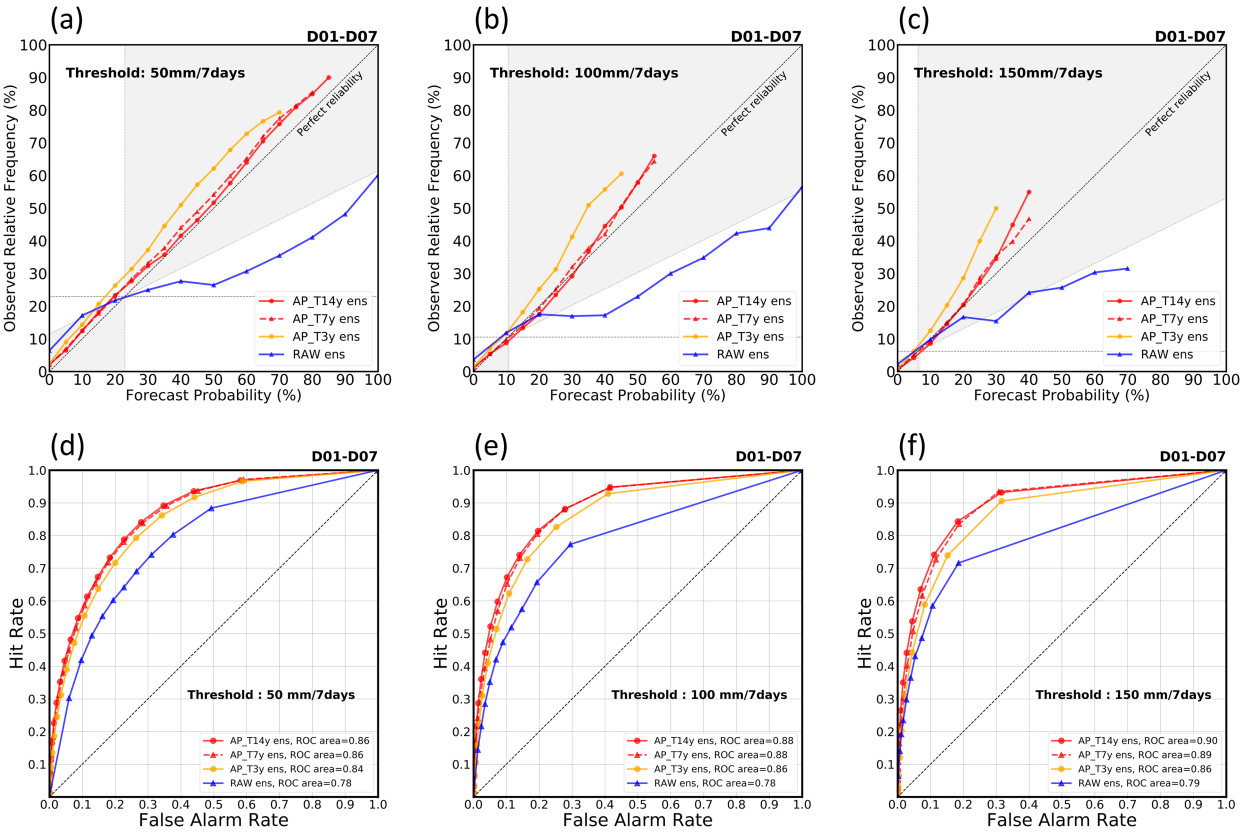

FIG. 15. (top) Reliability diagram and (bottom) ROC for REM (blue curve) and APPM (red curves) at (a),(d) 50, (b),(e) 100, and (c),(f) 150 mm week^{-1} for 1–7-day lead time. Evaluation is conducted over an 8-yr sample period from 2013 to 2020. For the three APPM curves, the orange solid curve represents AP-T3y and red dashed and solid curves represent the results from the experiment AP-T7y and AP-T14y, respectively.

The calibrated single (deterministic) forecast (APPM) had a lower MAE and explained approximately 3–8-fold more variance in the observational precipitation analysis in all seasons and lead times than did the REM forecast. A detailed evaluation showed that the RAW ensemble forecasts were underdispersive with an obvious wet bias, while the AP ensemble spread well represented forecast uncertainty with bias removed on a fine (1 km × 1 km) output grid. The improved AP ensemble spread contributed to better reliability in the AP-based PQPF. Discrimination was also improved, resulting in higher EV in the AP-based PQPF for a wider range of users, compared with the RAW forecasts. In summary, precipitation forecasts downscaled and calibrated by APPM substantially improved forecast quality and value.

The methods, AP and APPM, have been applied to operational forecasting (Chang et al. 2022) and have also been used to generate various customized precipitation products for agriculture and stock farming (https://agr.cwa.gov.tw/NAGR/) and water resources management sectors (https://qpeplus.cwa.gov.tw/WRA/) in Taiwan. One of the customized forecast products using the AP method is a probabilistic forecast of consecutive clear days, which has been employed in the stock farming industry for agricultural planning in the production of dry hay (Chou et al. 2023). Importantly, APPM precipitation forecasts have also been used as input for hydrological models predicting reservoir inflow for water resource management (Chang et al. 2023; Yang et al. 2023).

Acknowledgments. The authors greatly appreciate the valuable comments from Dr. Jeffrey Duda at the Global Systems Laboratory (GSL) of National Oceanic and Atmospheric Administration (NOAA). Special thanks are extended to the Central Weather Administration (CWA) for providing computer resources. This work was financially supported by the Ministry of Science and Technology (MOST) of Taiwan (104-2923-M-008-003-MY5 and 111-2111-M-008-021). HLC and PLL received a Grant from the MOST (110-2111-M-008-021).

Data availability statement. The reforecast data from the NOAA/NWS/EMC GEFS collected in the SubX are available through the portal https://downloads.psl.noaa.gov/Projects/NMME/SubX/EMC-GEFS/. The rain gauge precipitation data used to produce gridded precipitation analysis are available at https://codis.cwa.gov.tw/StationData.

APPENDIX A

Gridded Precipitation Analysis Based on Rain Gauge Data

For climatology analysis, the CWA generated a series of high-resolution station-based analyses using various kriging methods for a set of meteorological variables. The specific name assigned to these analyses is the Taiwan station-based analysis, referred to as TaiSA v1.0. It encompasses observational analysis data spanning from 1998 to the preceding year and is focused on six key meteorological variables: temperature, precipitation, pressure, relative humidity, specific humidity, and dewpoint temperature. The data are available in two resolutions, 2.5 and 1 km, respectively. To ensure homogeneity in climatology analysis, the number of meteorological stations (=319) used for observational analysis remains consistent every year.

However, the number of rain gauge stations has rapidly increased from 351 to 608 between 1999 and 2000 (see Fig. A1 and Table A1). To obtain well-calibrated high-resolution precipitation forecasts, it is crucial to have a set of precipitation ground truth reflecting real precipitation scenarios. Consequently, we generated another precipitation analysis based on all rain gauge observations over the land area of Taiwan for the same period, utilizing the same method as TaiSA v1.0. This dataset will be named TaiSA v1.1, and the primary difference in precipitation analysis between TaiSA v1.0 and v1.1 lies in the number of rain gauge stations used for precipitation analysis.

The average spacing of the observation rain gauge in 2020 was approximately 3.8 km for plain areas (terrain height $<$ 500 m) and approximately 5.5 km for mountainous areas (terrain height $\geq$ 500 m) over the land area of Taiwan. For plain areas, the minimum spacing between stations was 1 km, with only a few exceptions where the distance between stations was slightly less than 1 km. In addition, station density was significantly lower in mountainous areas than in plain areas. The optimal resolution for observational precipitation analysis needed to be determined in order to capture the true precipitation characteristics.

Figure A2 shows an example of the observational precipitation analysis with different resolutions from a typhoon case. If a resolution of 1 km (close to the minimum station spacing; Fig. A2a) is adopted for observational precipitation analysis, it does not introduce unrealistic precipitation characteristics in the mountainous areas. Instead, it just presents the precipitation characteristics on a spatial scale that can be resolved by the density of rain gauges in the mountainous areas. However, if we consider that the spacing of gauge stations in the mountainous areas cannot support a 1-km resolution of precipitation analysis, and opt for a coarser resolution (e.g., a 20-km resolution; Fig. A2d), it would sacrifice the ability to capture detailed precipitation characteristics in areas with high-density stations. Considering these factors, we generated a 1-km resolution of observational precipitation analysis as the ground truth required for the AP in this study.

We also used estimation error (estimated value − observed value) to quantify how confident we can be in the analysis results. Because observed values are only available at stations, the estimation error must be calculated at the station locations. Therefore, we conducted a shadowing experiment to calculate the estimation error at all stations. We removed

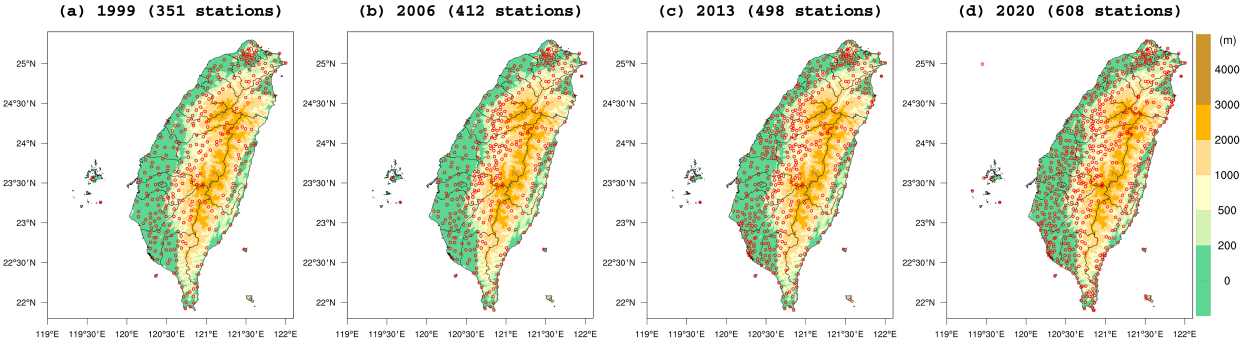

FIG. A1. Spatial distribution of rain gauges over the land area of Taiwan in different years.

TABLE A1. Typical spacing of the observation rain gauges in plain and mountainous areas of Taiwan in different years.

		1999	2006	2013	2020	Avg
Plain area (terrain height $<$ 500 m)	No. of gauges	256	276	357	436	331
	Avg spacing of gauges	5.0	4.8	4.2	3.8	4.4
Mountainous area (terrain height $\geq$ 500 m)	No. of gauges	95	136	141	172	136
	Avg spacing of gauges	7.4	6.2	6.1	5.5	6.3

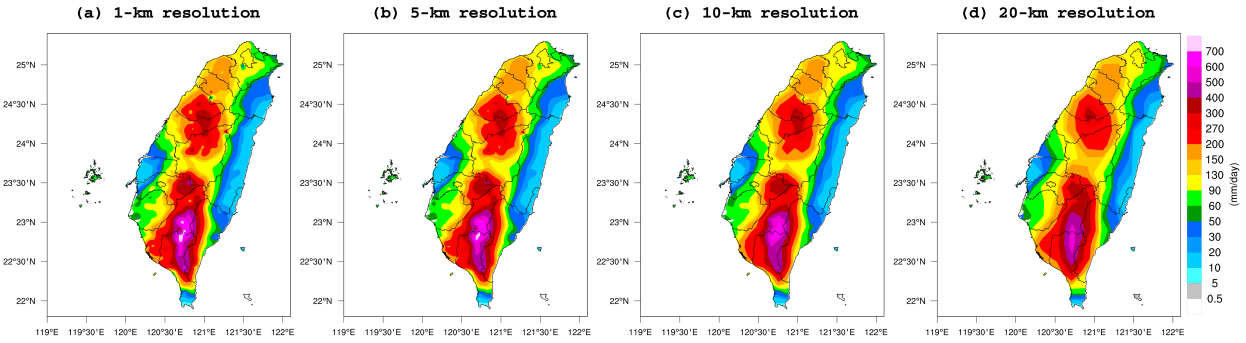

FIG. A2. (a) 1-, (b) 5-, (c) 10-, and (d) 20-km resolutions of observational analyses for daily precipitation ending at 0000 local standard time (LST) 8 Aug 2021 (Typhoon Lupit).

TABLE A2. Spatial (all grid points across the land area of Taiwan) and temporal (52 weeks) average estimation error for different years, including MAE and ME.

	1999	2006	2013	2020	avg
MAE (mm week^{-1})	12.1	12.5	10.5	9.6	11.2
ME (mm week^{-1})	−0.7	−0.7	−0.3	−0.2	−0.5

(shadowed) the observed precipitation value (Ro) at station A and used the remaining stations to estimate the precipitation value at station A (Re). We could then calculate the estimation error (Err) for station A (Err = Re − Ro). The average estimation error is shown in Table A2, and the spatial distribution of the average estimation error in different years is shown in Fig. A3.

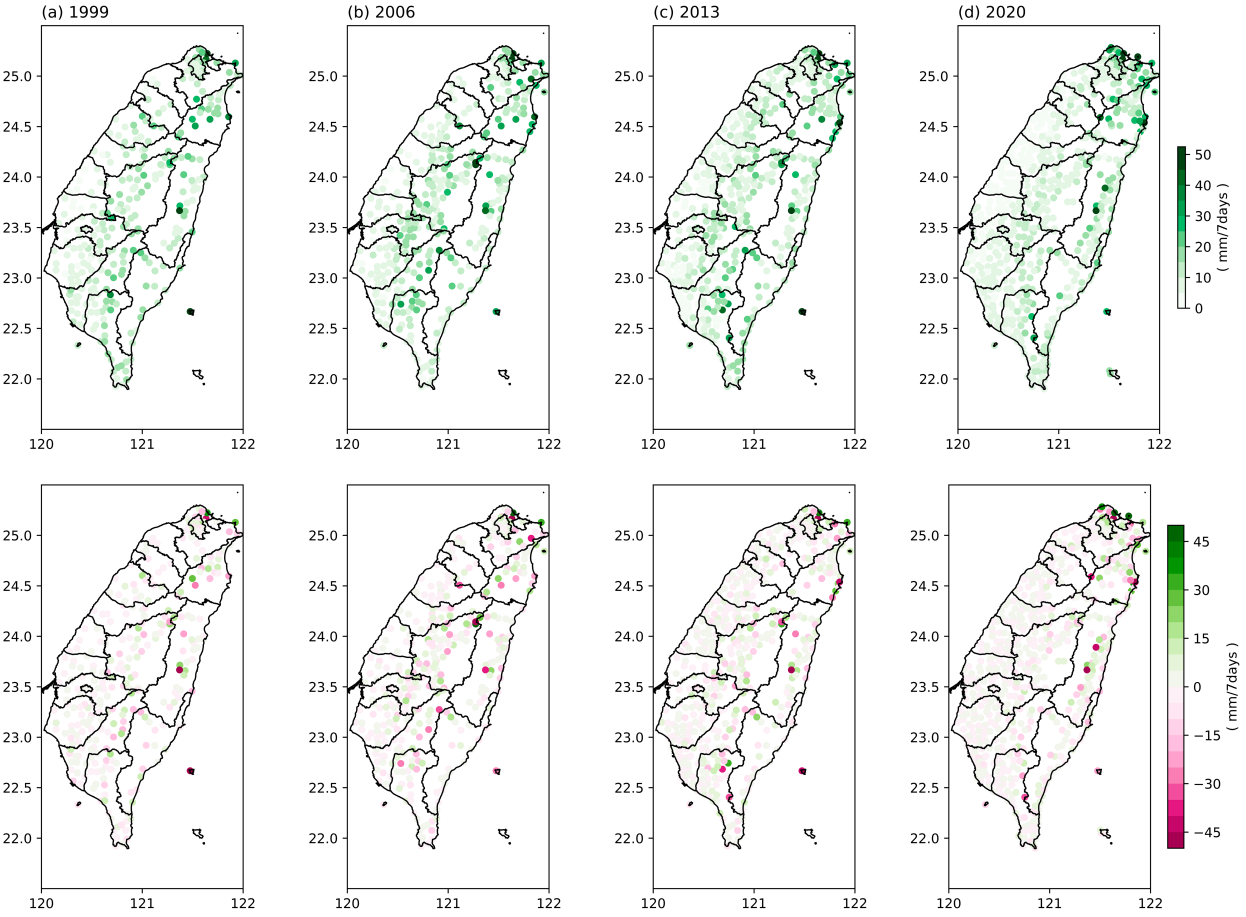

FIG. A3. Spatial distribution of MAE and ME in different years.

APPENDIX B

Large-Scale Circulation Indices for Pattern Matching

The large-scale indices used as an alternative for pattern matching for the winter half year, mei-yu, and summer are shown in Fig. B1 and described below.

a. Large-scale circulation indices for the winter half year (October–April)

The two selected indices used for pattern matching were TWRI1000 and TWRI850 (Figs. B1a,d). TWRI1000 = $V_{1000\text{hPa,A}} - V_{1000\text{hPa,B}}$, where $V_{1000\text{hPa,A}}$ and $V_{1000\text{hPa,B}}$ are the average meridional wind at 1000 hPa over areas A (125°–132.5°E, 12.5°–17.5°N) and B (117.5°–125°E, 25°–30°N). TWRI1000 reflects the low-level meridional convergence over Taiwan. Positive TWRI1000 values indicate that Taiwan is in a positive vorticity environment and is usually influenced by frontal systems. TWRI850 = $V_{850\text{hPa,A}} + V_{850\text{hPa,B}}$, where $V_{850\text{hPa,A}}$ and $V_{850\text{hPa,B}}$ are the average meridional wind at 850 hPa over areas A (120°–130°E, 12.5°–27.5°N) and B (130°–140°E, 15°–30°N). TWRI850 reflects the intensity of the low-level southerly wind component near Taiwan. Positive TWRI850 values indicate large-scale southerly winds that facilitate moisture transfer to Taiwan.

b. Large-scale circulation indices for the mei-yu season (May–June)

The two selected indices used for pattern matching were Hshear and Vshear (Figs. B1b,e). These indices were also used as precipitation monitoring indices in the mei-yu season at the CWB. Hshear = $U_{850\text{hPa,A}} - U_{850\text{hPa,B}}$, where $U_{850\text{hPa,A}}$ and $U_{850\text{hPa,B}}$ are the average zonal winds at 850 hPa over areas A (115°–125°E, 17.5°–22.5°N) and B (110°–120°E, 25°–30°N). Hshear reflects the horizontal wind shear or low-level circulation over Taiwan. Positive Hshear values indicate that Taiwan is in a positive vorticity environment. Vshear = $U_{850\text{hPa,A}} - U_{200\text{hPa,A}}$, where $U_{850\text{hPa,A}}$ and $U_{200\text{hPa,A}}$ are the average zonal winds at 850 and 200 hPa over area A (110°–130°E, 12.5°–20°N). Vshear reflects the vertical wind shear. The transition of Vshear from negative to positive represents a change in large-scale circulation from winter to summer monsoon over Taiwan.

c. Large-scale circulation indices for the summer season (July–September)

The two selected indices used for pattern matching were Hshear and South China Sea summer monsoon (SCSSM; Figs. B1b,c). SCSSM = $V_{1000\text{hPa,A}}$, which is the average meridional wind at 1000 hPa over area A (107.5°–120°E, 7.5°–20°N). SCSSM was used to monitor the change of meridional wind over the South China Sea. Positive SCSSM values indicate large-scale southerly winds that facilitate moister transfer to Taiwan.

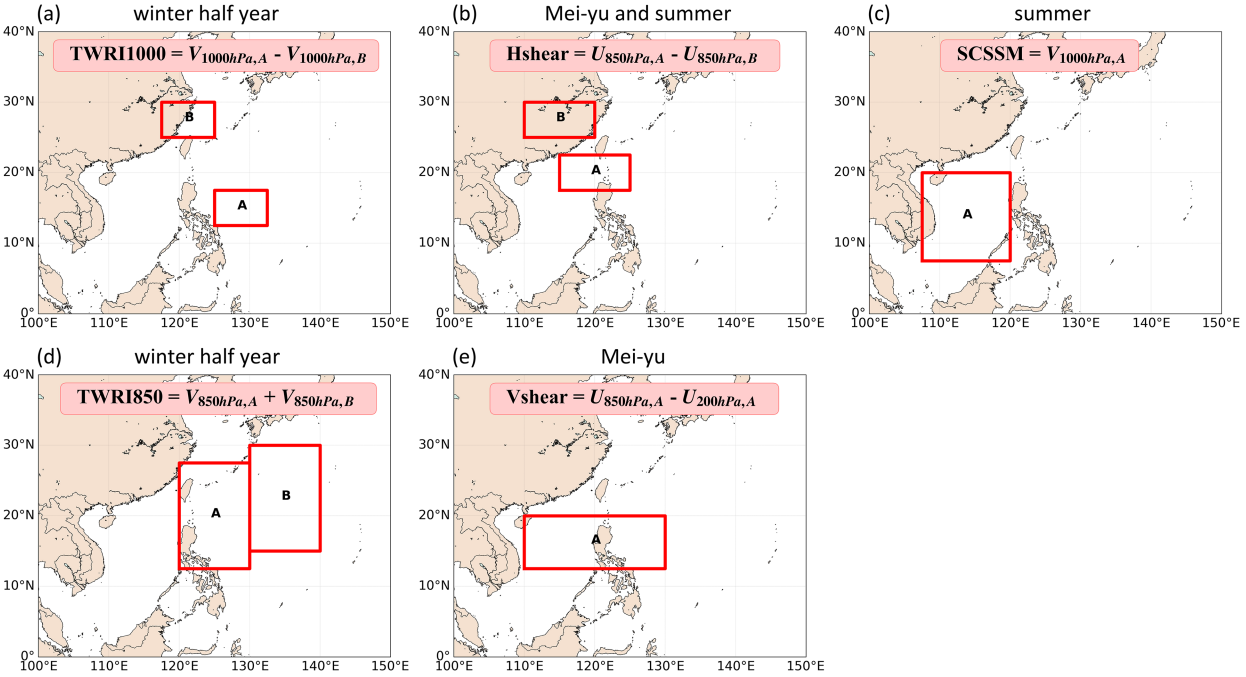

FIG. B1. (a)–(e) Large-scale circulation indices used for pattern matching in the winter half year in (a) and (d), mei-yu season in (b) and (e), and summer in (b) and (c). The definition of each index is shown in each plot, and red boxes in each plot display areas of the meteorological fields used for each index.

REFERENCES

Ali, A., T. Lebel, and A. Amani, 2005: Rainfall estimation in the Sahel. Part I: Error function. *J. Appl. Meteor.*, **44**, 1691–1706, https://doi.org/10.1175/JAM2304.1.

Ben Daoud, A., E. Sauquet, G. Bontron, C. Obled, and M. Lang, 2016: Daily quantitative precipitation forecasts based on the analogue method: Improvements and application to a French large river basin. *Atmos. Res.*, **169**, 147–159, https://doi.org/10.1016/j.atmosres.2015.09.015.

Boer, G. J., 2003: Predictability as a function of scale. *Atmos.–Ocean*, **41**, 203–215, https://doi.org/10.3137/ao.410302.

Chang, H.-L., H. Yuan, and P.-L. Lin, 2012: Short-range (0–12 h) PQPFs from time-lagged multimodel ensembles using LAPS. *Mon. Wea. Rev.*, **140**, 1496–1516, https://doi.org/10.1175/MWR-D-11-00085.1.

——, S.-C. Yang, H. Yuan, P.-L. Lin, and Y.-C. Liou, 2015: Analysis of relative operating characteristic and economic value using the LAPS ensemble prediction system in Taiwan area. *Mon. Wea. Rev.*, **143**, 1833–1848, https://doi.org/10.1175/MWR-D-14-00189.1.

——, B. G. Brown, P.-S. Chu, Y.-C. Liou, and W.-H. Wang, 2017: Nowcast guidance of afternoon convection initiation for Taiwan. *Wea. Forecasting*, **32**, 1801–1817, https://doi.org/10.1175/WAF-D-16-0224.1.

——, Z. Toth, S.-C. Chou, C.-Y. Feng, H.-F. Lin, and Y.-J. Chen, 2022: Statistical post-processing of 1–14 day precipitation forecasts for Taiwan. *Proc. 18th Annual Meeting of the Asia Oceania Geosciences Society (AOGS2021 eBook Extended Abstract Volume)*, Singapore, World Scientific Publishing Company, 25–27, https://doi.org/10.1142/9789811260100_0009.

——, T.-C. Yang, and J.-S. Hong, 2023: Extended-range reservoir inflow forecasting based on calibrated and downscaled rainfall forecasts. *19th Annual Meeting of the Asia Oceania Geosciences Society (AOGS2022 eBook Extended Abstract Volume)*, Singapore, World Scientific Publishing Company, 38–40, https://doi.org/10.1142/9789811275449_0013.

Chou, S.-C., H.-L. Chang, C.-Y. Feng, H.-F. Lin, and P.-L. Lin, 2023: Evaluation of probabilistic forecasts of consecutive days without measurable rainfall over Taiwan. *19th Annual Meeting of the Asia Oceania Geosciences Society (AOGS2022 eBook Extended Abstract Volume)*, Singapore, World Scientific Publishing Company, 35–37, https://doi.org/10.1142/9789811275449_0012.

Chu, P.-S., 2002: Large-scale circulation features associated with decadal variations of tropical cyclone activity over the central North Pacific. *J. Climate*, **15**, 2678–2689, https://doi.org/10.1175/1520-0442(2002)015<2678:LSCFAW>2.0.CO;2.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**, 618–641, https://doi.org/10.1175/2009JHM1006.1.

Horton, P., C. Obled, and M. Jaboyedoff, 2017: The analogue method for precipitation prediction: Finding better analogue situations at a sub-daily time step. *Hydrol. Earth Syst. Sci.*, **21**, 3307–3323, https://doi.org/10.5194/hess-21-3307-2017.

——, M. Jaboyedoff, and C. Obled, 2018: Using genetic algorithms to optimize the analogue method for precipitation prediction in the Swiss Alps. *J. Hydrol.*, **556**, 1220–1231, https://doi.org/10.1016/j.jhydrol.2017.04.017.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, https://doi.org/10.1016/0169-2070(86)90048-8.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 254 pp.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.*, **115**, 349–356, https://doi.org/10.1175/1520-0493(1987)115<0349:FFS>2.0.CO;2.

Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 222 pp.

Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *WIREs Water*, **4**, e1246, https://doi.org/10.1002/wat2.1246.

Lu, C., H. Yuan, B. Schwartz, and S. Benjamin, 2007: Short-range forecast using time-lagged ensembles. *Wea. Forecasting*, **22**, 580–595.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate*, **26**, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2.

Mass, C. F., 2003: IFPS and the future of the National Weather Service. *Wea. Forecasting*, **18**, 75–79, https://doi.org/10.1175/1520-0434(2003)018<0075:IATFOT>2.0.CO;2.

Mendoza, P. A., B. Rajagopalan, M. P. Clark, K. Ikeda, and R. M. Rasmussen, 2015: Statistical postprocessing of high-resolution regional climate model output. *Mon. Wea. Rev.*, **143**, 1533–1553, https://doi.org/10.1175/MWR-D-14-00159.1.

Murphy, A. H., 1977: The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816, https://doi.org/10.1175/1520-0493(1977)105%3C0803:TVOCCA%3E2.0.CO;2.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment.

Bull. Amer. Meteor. Soc., **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Su, Y.-J., J.-S. Hong, and C.-H. Li, 2014: The characteristics of the probability matched mean QPF for 2014 Meiyu Season (in Chinese with an English abstract). *Atmos. Sci.*, **22**, 113–134.

Toth, Z., 1989: Long-range weather forecasting using an analog approach. *J. Climate*, **2**, 594–607, https://doi.org/10.1175/1520-0442(1989)002<0594:LRWFUA>2.0.CO;2.

——, Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 463–477, https://doi.org/10.1175/1520-0434(2001)016<0463:TUOETI>2.0.CO;2.

——, O. Talagrand, and Y. Zhu, 2006: The attributes of forecast systems: A general framework for the evaluation and calibration of weather forecasts. *Predictability of Weather and Climate: From Theory to Practice*, T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 584–595.

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.*, **117**, 2230–2247, https://doi.org/10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2.

Vannitsem, S., D. S. Wilks, and J. W. Messner, 2019: *Statistical Postprocessing of Ensemble Forecasts*. 1st ed. Academic Press, 362 pp.

——, and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz'96 setting. *Meteor. Appl.*, **13**, 243–256, https://doi.org/10.1017/S1350482706002192.

——, 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, https://doi.org/10.1175/MWR3402.1.

Yang, T.-C., M.-J. Kung, L.-C. Pi, H.-L. Chang, J.-S. Hong, and P.-S. Yu, 2023: Integration of extended-range and long-term rainfall forecasts towards reservoir inflow forecasting. *J. Taiwan Agric. Eng.*, **69**, 28–41.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H.-M. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network *Wea. Forecasting*, **22**, 1287–1303, https://doi.org/10.1175/2007WAF2006114.1.

——, J. A. McGinley, P. J. Schultz, C. J. Anderson, and C. Lu, 2008: Short-range precipitation forecasts from time-lagged multimodel ensembles during the HMT-West-2006 campaign *J. Hydrometeor.*, **9**, 477–491, https://doi.org/10.1175/2007JHM879.1.

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

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability. *Adv. Atmos. Sci.*, **22**, 781–788, https://doi.org/10.1007/BF02918678.

——, and Y. Luo, 2015: Precipitation calibration based on the frequency-matching method. *Wea. Forecasting*, **30**, 1109–1124, https://doi.org/10.1175/WAF-D-13-00049.1.

——, Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–84, https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate*, **12**, 2474–2489, https://doi.org/10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.