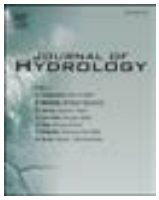
Journal of Hydrology 660 (2025) 133484

Contents lists available at ScienceDirect

Journal of Hydrology

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



# Ensemble post-processing of sub-seasonal to seasonal precipitation forecasts based on a novel probabilistic double machine learning method

Shengsheng Zhan [a], Aizhong Ye [a,*], Lingyun Wu [a], Chenguang Zhao [a]

[a] State Key Laboratory of Earth Surface Processes and Disaster Risk Reduction, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

*Keywords:*
Sub-seasonal to seasonal
Precipitation forecasts
Post-processing
Probabilistic double machine learning

ABSTRACT

Subseasonal-to-seasonal (S2S) precipitation forecasting is crucial for hydrological modeling; however, its accuracy often falls short of the requirements for hydrological forecasts, necessitating post-processing. A novel improved version of the Double Machine Learning (DML) method, termed Probabilistic Double Machine Learning (PDML), is proposed for ensemble post-processing of S2S forecasts. The new PDML method extends the classifier from binary classification to multi-class classification, improves the regressor from single-value output to probability distribution output, and combines the classifier and regressor based on total probability theorem. PDML not only quantifies uncertainty through ensemble output but also provides additional consideration for extreme precipitation events in the classification and regression progress. Various machine learning methods are compared within the PDML framework, including the state-of-the-art Kolmogorov-Arnold Networks. The results indicate that deep learning models based on Recurrent Neural Networks (RNN) and the U-NET architecture perform the best within the PDML framework. It achieves post-processing of S2S forecasts across different timescales and outperforms the statistical Ensemble Pre-Processor (EPP) method. On average, it improves the original forecast's correlation coefficient, critical success index, and root mean square error by 85.8 %, 294.6 %, and 45.3 %, respectively, and achieves an 8.6 % improvement on the continuous ranked probability score compared to EPP. The results demonstrate that PDML can effectively perform ensemble post-processing of precipitation forecasts across different timescales, quantify uncertainty, and facilitate further hydrological modeling.

## 1. Introduction

The subseasonal-to-seasonal (S2S) forecasts, mainly spanning timescales from two weeks to one season, have significant potential to augment existing weather and climate services and products. This forecasting framework demonstrates considerable applicability across a range of sectors, including public health, agriculture, water resource management, disaster mitigation, renewable energy and utilities, and emergency management and response (White et al., 2022). Besides, S2S forecasts present a valuable opportunity for numerous industries, facilitating the ability to engage in systematic planning within this novel time horizon (White et al., 2017). Among these, precipitation is the most important atmospheric component of the hydrologic cycle and perhaps the most important and primary input to most hydrological models that are employed for planning, design, and operation of water resource projects (Duan et al., 2019). Precipitation forecasts is crucial for hydrological research, including flood and drought prediction, as well as hydrological system modeling (Chen et al., 2024; Luo et al., 2024; Ye et al., 2017).

The S2S forecasting is influenced both by atmospheric initial conditions and by slowly evolving boundary conditions, distinguishing it from short-term numerical weather prediction (NWP) and from long-term global circulation models (GCM) (L. Zhang et al., 2023). Due to the unique nature of the S2S timescale, the memory of atmospheric initial conditions is largely lost, while the variability of the ocean has yet to exert a strong influence. This results in a lack of predictability for S2S forecasts, which have long been in a forecasting gap, often referred to as a "predictability desert" (Vitart, 2014; Vitart et al., 2017). Currently, the accuracy of S2S precipitation forecasts is insufficient to meet practical demands as verification studies have identified two major shortcomings: a rapid increase in bias as the forecast lead time increases, and inadequate capability to predict extreme precipitation events (Liu et al., 2023; Rivoire et al., 2023; Lujun Zhang et al., 2021).

The S2S precipitation forecasts with systematic bias and dispersion

* Corresponding author.
*E-mail address:* azye@bnu.edu.cn (A. Ye).

https://doi.org/10.1016/j.jhydrol.2025.133484
Received 4 February 2025; Received in revised form 24 March 2025; Accepted 6 May 2025
Available online 8 May 2025

errors must undergo post-processing to improve accuracy before being applied to hydrological simulations (Li et al., 2017; Manzanas et al., 2019). Kolachian and Saghafian (2019) tested two post-processing methods, quantile mapping (QM) and Bayesian model averaging (BMA), for S2S precipitation forecasts, and found that the BMA method outperformed QM. Li et al. (2023) adjusted the QM method for the S2S timescale and proposed an improved version called quantile mapping of matching precipitation threshold by time series (MPTT-QM) method. Specq and Batté (2020) developed a statistical-dynamical scheme within a Bayesian framework, which demonstrated more reliable calibration results compared to the original forecast for large-scale climate phenomena such as El Niño-Southern Oscillation (ENSO) and Madden Julian Oscillation (MJO). Huang et al. (2022) used a seven-parameter Bernoulli-Gamma-Gaussian model to calibrate S2S precipitation forecasts, improving reliability and yielding forecast skill for daily and accumulated precipitation. In addition to the aforementioned statistical methods, recent years have seen machine learning (ML) and its subfield, deep learning (DL) methods, demonstrate significant potential within precipitation forecast post-processing. ML methods employed in precipitation forecast post-processing offer several key advantages: 1) Flexibility in modeling nonlinear relationships between input variables and output predictions, enabling the extraction of meaningful information from multivariate datasets; 2) No need for stringent assumptions regarding the underlying data distribution; 3) A well-established research ecosystem and established frameworks that facilitate practical implementation (Oliveira et al., 2023; Zhang and Ye, 2021). Numerous studies have demonstrated the feasibility of various machine learning algorithms in precipitation forecast post-processing, such as random forests (RF) (Herman and Schumacher, 2018; Mao and Sorteberg, 2020; Vitart et al., 2022), support vector machines (SVM) (Ortiz-García et al., 2014; Yin et al., 2023, 2022), Long Short-Term Memory Networks (LSTM) (Chen et al., 2022; Li et al., 2021; Ni et al., 2020), Convolutional Neural Networks (CNN) (Li et al., 2022; Lyu et al., 2024, 2023; Vitart et al., 2022; Weyn et al., 2020), and other related works.

In addition to improvements in algorithms, a new machine learning framework, Double Machine Learning (DML), has been gaining increasing popularity in the post-processing of precipitation data in recent years (Ling Zhang et al., 2021). The DML approach modifies the conventional single-regression framework in machine learning-based post-processing by initially employing a classifier to categorize the precipitation into its respective levels, prior to the application of a regressor. Subsequently, the outcomes of both the classifier and the regressor are integrated, typically with the classifier's results superseding those of the regressor, to produce the final forecast. This methodology has been validated in the calibration of various precipitation products, demonstrating its feasibility (Lei et al., 2022; Lyu and Yong, 2024; Senocak et al., 2023; Xiao et al., 2022). Several studies have confirmed that DML, compared to the traditional single-regression ML approach, can lead to further performance improvements (Kossieris et al., 2024; Ling Zhang et al., 2021). However, this framework remains in infancy and still presents certain limitations, necessitating further exploration and development.

There are two main shortcomings in the existing research on DML, in our view. First, most studies have limited the classifier to a binary classification, i.e., determining whether precipitation occurs or not (Lei et al., 2022; Lyu and Yong, 2024; Xiao et al., 2022; Ling Zhang et al., 2021). In Senocak et al.(2023), an attempt was made to extend the binary classification to a multi-class classification, while the method used involved passing the classifier's output as an additional feature to the regressor, thereby overlooking the potential negative impact of classification errors. Second, to the best of our knowledge, almost all DML studies have opted for a single-value forecast output. While deterministic results are simpler, they fail to capture the inherent uncertainty associated with the predictions (Klotz et al., 2022; Wang, 2001; Zhang et al., 2022). Therefore, we believe it is necessary to design a probabilistic post-processing method for the DML framework, which not only quantifies uncertainty but also provides more comprehensive reference information for decision-makers (Scheuerer et al., 2020; Tao et al., 2014; Y. Zhang et al., 2023).

In this study, we propose a new Probabilistic Double Machine Learning (PDML) method for post-processing S2S precipitation forecasts. Compared to DML, the new PDML method extends the classifier from binary classification to multi-class classification, improves the regressor from single-value output to probability distribution output, and combines the classifier and regressor based on total probability theorem. Under the new framework, uncertainty can be quantified through the ensemble forecasts output, and the ability to identify extreme precipitation is enhanced by distinguishing extreme precipitation mappings. Within the PDML, we compare five different machine learning algorithms, including four commonly used algorithms and a recent developed Kolmogorov-Arnold Networks (KAN) algorithm, with an improved version of statistical method Ensemble Pre-Processor (EPP) also included as a benchmark for comparison (Li et al., 2019; Liu et al., 2024). Moreover, an interpretability analysis of the PDML model was conducted to understand the impact of features on the output parameters, which has not been achieved in DML studies. The structure of the subsequent sections is as follows: Section 2 introduces the data used in this study; Section 3 presents the post-processing methods and evaluation metrics employed, as well as the interpretability analysis methods; Section 4 displays the research results and provides a discussion; Section 5 concludes the paper.

## 2. Data

### 2.1. Study area

The study area is defined as the mainland China, encompassing the longitudinal range of 73° E to 135°E and the latitudinal range of 18° N to 53° N. The annual average precipitation in mainland China is 607 mm/a. Most of mainland China is controlled by a monsoon climate, with precipitation showing significant seasonal variation.. The spatial distribution of annual average precipitation and the monthly average precipitation process curve in mainland China are shown in Fig. 1. It can be observed that precipitation in mainland China exhibits significant spatial and temporal distribution disparities, which imposes higher demands for accurate precipitation forecasts.

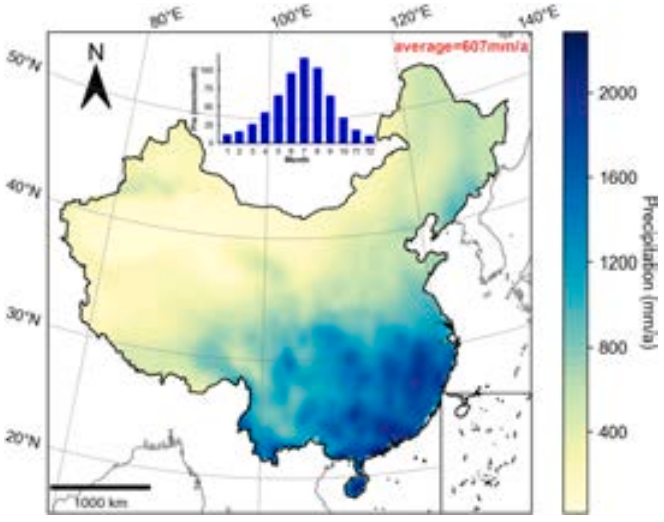



**Fig. 1.** Study area map with distribution of the annual precipitation in mainland China (mm/a) with the bar chart showing the monthly average precipitation in mainland China (mm/month).

### 2.2. Forecast data

The forecast data used in this study is derived from the CAS-FGOALS-f2-V1.3 forecasting model developed by the Institute of Atmospheric Physics, Chinese Academy of Sciences (IAP-CAS). This model was incorporated into the Sub-seasonal to Seasonal Prediction Project on January 5, 2021, and has since been providing operational forecasts and reporting data (Vitart et al., 2017).

We utilized the reforecast data from 1999 to 2018, ensuring that the model version remained unchanged throughout this period. In addition to the precipitation (PRE) control forecast data, other auxiliary variables were also incorporated into the ML approach, including the geopotential height (GH), specific humidity (SH), and temperature (T) control forecast at three different pressure levels (200 hPa, 500 hPa, and 850 hPa), which means that a total of 10 features are used (PRE, GH200, GH500, GH850, T200, T50, T850, SH200, SH500, SH850). These auxiliary variables have been demonstrated to be effective in previous studies (Lyu et al., 2023). The forecast data has a spatial resolution of 1.0° × 1.0° and a forecast period of up to 65 days with a reforecast frequency of daily. To comprehensively compare the performance of the PDML algorithm, we tested eight different lead periods, including four daily precipitation scales (forecast periods of day 1, day 2, day 3, and day 4) and four accumulated precipitation scales (forecast periods of 1–7 days, 8–14 days, 15–30 days, and 31–60 days, averaged).

The reasons for selecting this model are as follows: 1) The IAP-CAS model provides daily reforecast data, which significantly increases the available data volume by more than three times compared to the ECMWF reforecasts, which typically has a biweekly frequency. In comparison to the NCEP reforecasts, which also provides daily data but covers only 11 years, the IAP-CAS reforecasts offer a 20-year dataset. Although the focus of this study is not on the sensitivity analysis of machine learning algorithm performance with respect to sample size, we still chose the forecast model with the largest sample size to compare different machine learning algorithms. A larger sample size is generally expected to yield more robust results and greater performance improvements in data-dependent post-processing methods like ML (Fassnacht et al., 2014; Moghaddam et al., 2020); 2) The IAP-CAS model has maintained a consistent version throughout the complete fixed reporting period, ensuring a stable forecast-to-observation mapping relationship.

### 2.3. Observation data

The precipitation product CN05.1, generated through optimal interpolation based on daily observation data from over 2,400 national meteorological stations provided by the National Meteorological Information Center, was selected as the reference data (Xu et al., 2009). This dataset is extensively utilized in precipitation research focused on mainland China, including applications such as trend analysis, model validation, and forecast post-processing (Liu et al., 2023; Lyu et al., 2024; Wu et al., 2017).

The original spatial resolution of the CN05.1 data is 0.25° × 0.25°, which is upscaled to 1.0° × 1.0° using spatial averaging to match the resolution of the forecast data. After standardizing the spatial scale, there are a total of 1,068 grid points across mainland China. The temporal resolution is daily, and we aligned it with the different lead time of the forecast data.

## 3. Methods

### 3.1. Probabilistic Double Machine learning structure

The Probabilistic Double Machine Learning (PDML) method aims to combine machine learning classifiers and regressors in the form of probability distributions to correct the raw forecasts, thereby generating precipitation forecasts with higher accuracy and lower error. The technical approach of this method is illustrated in Fig. 2.

The three main improvements of PDML over DML are: 1) It replaces binary classification with multi-class classification and generates probabilities instead of deterministic results, by doing so can extreme precipitation events be considered; 2) The regressor uses probabilistic output to obtain a statistical distribution rather than a single value output, which allows to quantify the uncertainty; 3) PDML integrates classifiers and regressors in a probabilistic framework, consistent with the mathematical principles of total probability theorem, whereas DML merely overlays the classifier results with the regressor results, making the interpretation more challenging. Next, we will provide an introduction to the framework and algorithm details of PDML.

First, we align the forecast data, including precipitation forecasts and other meteorological forecast data, with the observation data to the same scale. Then, for the 20 years of available data, we divide it into three groups: a 12-year training set (1999–2010), a 4-year validation set (2011–2014), and a 4-year test set (2015–2018). The training set data is used to train the model, the validation set data is used to adjust the hyperparameters, and the test set data is used for final result evaluation.

Next, based on the observation data from the training set, we determined two thresholds, $T_{01}$ and $T_{12}$ to classify precipitation into three levels ($level_0$, $level_1$, $level_2$). The thresholds are set individually for each grid. The threshold between $level_0$ and $level_1$ is the zero precipitation threshold, which is defined as follows: the daily precipitation data for each grid point in the corresponding forecast period is sorted in descending order, denoted as $\{a_1, a_2, \ldots, a_m\}$, then set a percentile value $P_a$ and $a_n$ is determined based on formula (1) (Ye et al., 2017). $T_{01}$ is taken as the smaller value between $a_n$ and 0.1, i.e., $min(0.1, a_n)$. Precipitation below $T_{01}$ is considered as an artifact generated by interpolation and is treated as no precipitation occurrence. The threshold between $level_1$ and $level_2$ is the threshold between normal precipitation and extreme precipitation, which is defined as the 90th percentile of all precipitation events greater than $T_{01}$ in the corresponding forecast period. This means that for each grid point and each different forecast period, two thresholds are determined separately.

$$P_a = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{m} a_i} \tag{1}$$

After processing all the data through Z-score standardization, the training set data is used to train the classifier and regressor separately. The purpose of the classifier is to identify the level of precipitation to which a given sample belongs. Unlike previous DML studies where the classifier uses deterministic output results, in PDML, we use the Softmax activation function in formula (2) to output the probabilities of a sample belonging to each level. The Softmax activation function is suitable for multi-class tasks and can map the output values to probability values that sum to 1, representing the probabilities of the classification result belonging to each category. Compared to deterministic classification results in DML, this approach not only applies to generating ensemble distributions but also reduces the cost of classification errors. The loss function used by the classifier is the cross-entropy, as shown in formula (3). After training, for a given sample, the classifier outputs the probabilities $p_0$, $p_1$, and $p_2$, which represent the likelihood of the sample belonging to the three categories: $level_0$, $level_1$, and $level_2$, respectively, given the ten input features.

$$Softmax(x_i) = \frac{exp(x_i)}{\sum_{j=0}^{2} exp(x_j)} \tag{2}$$

Where $[x_0, x_1, x_2]$ is the vector of raw outputs from the classifier, which can be transformed into a probability vector $[p_0, p_1, p_2]$ after passing through the Softmax activation function.

$$L(y, \hat{y}) = -\sum_{i=0}^{2} y_i \log(\hat{y}_i) \tag{3}$$

Where $[y_0, y_1, y_2]$ is the true label distribution, $y_i$ being either

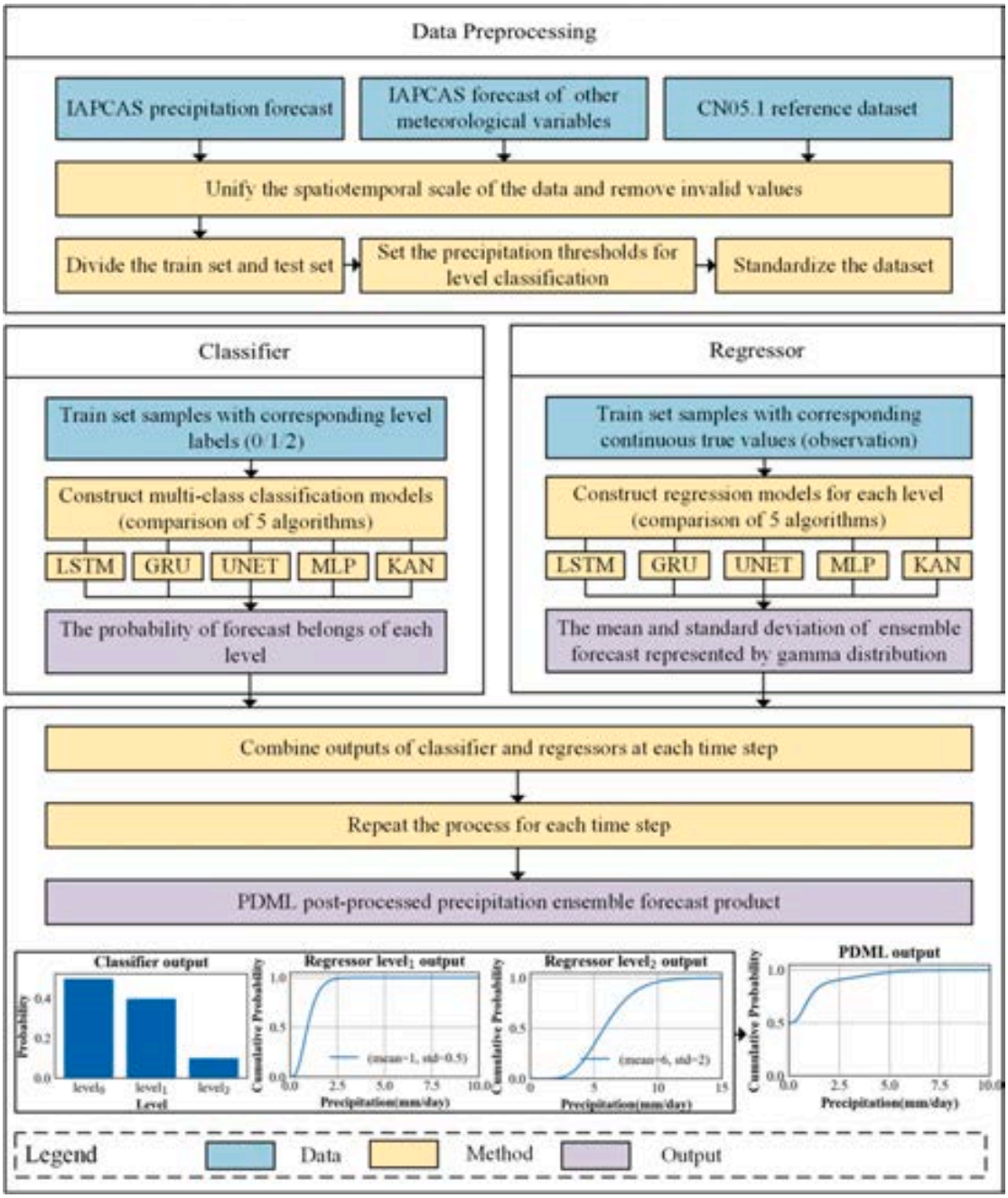



**Fig. 2.** Flowchart showing the structure of the PDML method.

0 (False) or 1 (True), and $[\hat{y}_0, \hat{y}_1, \hat{y}_2]$ is the predicted probability distribution by the model. For each sample, the smaller the loss, the more accurate the model's prediction of the true class.

We have also made improvements to the single-value regressor in traditional DML method. We train separate regressors for the two levels of precipitation ($level_0$ does not require a regressor since it represents no precipitation). These regressors characterize the distribution followed when precipitation belongs to the corresponding level. Considering that the Gamma distribution is suitable for representing the skewed and heavy-tailed characteristics of precipitation, we choose to output two parameters of the Gamma distribution to characterize the precipitation distribution (Martinez-Villalobos and Neelin, 2019). The regressors of each level will output the mean $\mu$ and standard deviation $\sigma$ of a Gamma distribution, which are then transformed into the shape and scale parameters of the Gamma distribution using the following formulas, thus generating a deterministic cumulative distribution function (CDF).

$$k = \frac{\mu^2}{\sigma^2}, \theta = \frac{\sigma^2}{\mu} \tag{4}$$

$$F_{k,\theta}(x) = \frac{1}{\Gamma(k)} \int_0^x t^{k-1} e^{-\frac{t}{\theta}} \frac{1}{\theta^k} dt \tag{5}$$

Where $k$ and $\theta$ are the shape parameter and scale parameter of the Gamma distribution, $\Gamma$ is the Gamma function, and $F$ is the CDF of the Gamma distribution.

The loss function used for training the regressor is the Continuous Ranked Probability Score (CRPS) loss function, whose closed-form expression for Gamma distribution is derived by Scheuerer and Hamill (2015). Since we train separate regressors for each level, only samples that belong to the corresponding level are included in the loss

calculation, as shown in formula (6). For example, when training the regressor for $level_1$, the sample losses for $level_0$ and $level_2$ are both zero and are therefore not included in the backpropagation.

$$L(F_{k,\theta}, y) = \begin{cases} y(2F_{k,\theta}(y)-1) - k\theta(2F_{k+1,\theta}(y)-1) - \dfrac{\theta}{B\left(\frac{1}{2}, k\right)} & if\ y \in level_i \\ 0 & else \end{cases} \tag{6}$$

Where $k$ and $\theta$ are the Gamma distribution parameters calculated from the model output using formula (4), $y$ is the observed precipitation, and $B$ is the Beta function.

In the combination of classifiers and regressors, PDML derives the integration formula based on the law of total probability. Suppose a precipitation event (including zero precipitation) can be classified into n categories based on intensity, denoted as $level_0, level_1, level_{n-1}$. Then the sum of the probabilities for all categories at any given time step equals unity, i.e:

$$\sum_{k=0}^{n-1} P(y \in level_k) = 1 \tag{7}$$

Then, based on the total probability theorem, the probability that a given precipitation event has a precipitation amount equal to $\hat{y}$ can be expressed as follows:

$$P(y=\hat{y}) = \sum_{k=0}^{n-1} P(y=\hat{y} | y \in level_k) \tag{8}$$

In our PDML case, a three-class classification is used, and the resulting calculation formula is:

$$P(y=\hat{y}) = \sum_{k=0}^{2} P(y=\hat{y} | y \in level_k) \tag{9}$$

Where $y \in level_0$, $y \in level_1$, and $y \in level_2$ represent the probabilities of the precipitation event belonging to zero-level precipitation (zero-precipitation), first-level precipitation, and second-level precipitation (extreme-precipitation), respectively. If we express the probability density in the form of a cumulative distribution, we obtain Equation (10):

$$F(x) = p_0 + p_1 F_{k_1,\theta_1}(x) + p_2 F_{k_2,\theta_2}(x) \tag{10}$$

Through the training process, we obtained a classifier that outputs the probability vector $[p_0, p_1, p_2]$, and two regressors, which output the Gamma distribution parameters and can calculate $k_1, \theta_1, k_2, \theta_2$. The final distribution obtained is a mixture of discrete (represented by $p_0$) and continuous (represented by $p_1, k_1, \theta_1, p_2, k_2, \theta_2$) distributions, as shown in the example provided at the end of Fig. 2.

The above describes the overall framework of the PDML post-processing method. In the actual training process of classifiers and regressors, different deep learning algorithms can be chosen. To systematically compare the performance differences of various neural networks within the PDML framework, we tested five different DL algorithms, including two types of recurrent neural networks: Long Short-Term Memory Networks (LSTM) and Gated Recurrent Unit (GRU); one type of convolutional neural networks: U-NET; and two simple feedforward neural networks: Kolmogorov-Arnold Networks (KAN) and Multi-Layer Perceptron (MLP). Under the PDML framework, we independently train the classification and regression models for each forecast period. In the supplementary materials Text. S1 and Fig. S1, we provided a concise overview of these five DL algorithms. Although we have selected only five common deep learning models for comparison within the PDML framework as benchmark experiments, PDML is a flexible architecture, and the choice of models is not fixed. Any supervised learning model or generative model that supports backpropagation can be applied within the PDML framework following the process outlined above.

Hyperparameters refer to the parameters that need to be set before training a ML or DL model, and they control the training process and structure of the model. Unlike model parameters (such as weights and biases), hyperparameters must be manually specified before training begins, whereas model parameters are learned automatically from the training data. The hyperparameter configurations of five machine learning algorithms are provided in the supplementary materials Table S1. All other hyperparameters were set to default values. Since most models are independently trained on different grids, performing hyperparameter tuning for each grid individually would incur a substantial computational cost. Therefore, hyperparameter tuning was conducted using a grid search method applied to a few randomly selected grid points on lead day 1 based on experience, and then applied to all grid points. The five machine learning methods are all implemented within the PDML framework. For the sake of brevity, we will omit the "PDML-" prefix in the subsequent discussion and refer to the machine learning algorithms by their respective architectures (e.g., GRU, LSTM).

### 3.2. Statistical benchmark: CMLE-EPP

Given that the original DML method can only generate single-value forecasts and lacks the ability to produce ensembles, it is not considered as a benchmark in this study. In response to the call of Xu et al (2024), we contend that when conducting research on machine learning-based post-processing methods for precipitation forecasts, it is crucial to select more advanced and state-of-the-art statistical methods as benchmarks, rather than opting for methods that are relatively easy to outperform (such as QM). Therefore, we have decided to use an improved Ensemble Pre-Processor (EPP), referred to as the Ensemble Pre-Processor by applying maximum likelihood estimation for censored data (CMLE-EPP), as the statistical benchmark (Li et al., 2019). The EPP method, developed by the U.S. National Weather Service, has shown substantial improvements in precipitation forecasting post-processing, and the enhanced version, CMLE-EPP, further augments its performance (Huang et al., 2022; Li et al., 2019; Tao et al., 2014; Ye et al., 2017).

The EPP post-processing primarily involves the following steps: 1) Normal Quantile Transformation (NQT) applied to both the forecasts and observations; 2) Estimation of the joint distribution between the transformed observations and forecasts in the Normal space; 3) Computation of the conditional distribution of observations given the new forecasts, followed by the inverse NQT. The improvement in EPP-CMLE lies in using the correlation coefficient of the joint distribution (bivariate Gaussian distribution with censored data) estimated via the Maximum Likelihood Estimation in Normal space after the quantile transformation, as opposed to directly using the correlation coefficient between forecasts and observations in the original space. For the available 20 years of data, 16 years (1999–2014) are used for model training, with the final 4 years (2015–2018) reserved for testing, consistent with the PDML method. Similarly, we extract 1,000 ensemble members from the conditional distribution of the observations, as the output of this approach. We did not implement the Schaake shuffle step; rather, we directly utilized the ensemble forecasts for the evaluation, as the ordering of the ensemble members does not influence the assessment of the subsequent metrics. The specific steps of the CMLE-EPP algorithm are described in supplementary materials **Text S2**, with further details available in the original work by Li et al (2019). To maintain conciseness, the method will be denoted as EPP in the subsequent text.

### 3.3. Evaluation metrics

#### 3.3.1. Classification accuracy evaluation metrics

We first conduct a preliminary comparison of classification performance among the five PDML models using Accuracy (ACC) and Cross-Entropy Loss (CE Loss) as the evaluation metric. Given that our classification task involves three categories, the ACC is calculated as follows:

$$ACC = \frac{\sum_{i=1}^{n} I(\hat{y}_{label} = y_{label})}{n} \tag{11}$$

Where $\hat{y}_{label}$ is the predicted label for a precipitation event, $y_{label}$ is the true label of the precipitation event. $I$ denotes the indicator function such that $I(\hat{y}_{label} = y_{label}) = 1$ if condition $\hat{y}_{label} = y_{label}$ is satisfied, and 0 otherwise. $n$ is the total number of time steps.

The range of ACC is $[0, 1]$, where a larger value indicates more accurate predictions. However, it is critical to emphasize that Accuracy (ACC) serves only as a preliminary indicator of PDML model performance. Because the classifier must be integrated with the regressor through Equation (10) to generate the final postprocessed results, deficiencies in the classifier may be offset by a high-performing regressor, while classification accuracy could conversely be degraded by regressor-induced errors. Additionally, our use of probabilistic outputs (where $\hat{y}_{label} = argmax([p_0, p_1, p_2])$) rather than deterministic predictions effectively mitigates misclassification costs when two or more class probabilities exhibit negligible differences. Crucially, while a high-ACC model may intuitively suggest robust PDML performance, even models with lower ACC values might achieve compensatory gains through probabilistic uncertainty quantification and high-precision regression. Consequently, ACC provides only partial insight into the PDML's holistic efficacy.

The calculation of CE Loss is shown in formula (15). Unlike ACC, which only considers whether the final prediction is correct, CE Loss takes into account not only whether the classification is correct but also the confidence of the prediction (i.e., the model's probability output). It penalizes samples with high confidence but incorrect predictions. The range of CE Loss is $[0, +\infty)$, where a smaller value indicates more accurate predictions.

$$CELoss = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k-1} y_{ij}\log\left(p_{ij}\right) \tag{12}$$

Where $n$ is the number of samples, $k$ is the number of classes, $y_{ij}$ is the true label of sample $i$ for class $j$ (one-hot encoded) and $p_{ij}$is the predicted probability that sample $i$ belongs to class $j$.

### 3.3.2. Deterministic forecast evaluation metrics

Although ensemble forecasts typically capture more uncertainty information, decision-makers tend to prefer deterministic (single-point) predictions (Y. Zhang et al., 2023). Therefore, we first evaluated the deterministic post-processed forecast results (using the ensemble mean). The evaluation metrics include the temporal correlation coefficient (TCC), root mean square error (RMSE), and critical success index (CSI). The formulas for these three evaluation metrics are as follows:

$$TCC = \frac{\sum_{t=1}^{n}(y_t - \bar{y})(x_t - \bar{x})}{\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2 \sum_{t=1}^{n}(x_t - \bar{x})^2}} \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - x_t)^2} \tag{14}$$

$$CSI = \frac{H}{H + M + F} \tag{15}$$

Where $y_t$ is the observed value at time $t$ and $x_t$ is the forecast value at time $t$. $\bar{y}$ and $\bar{x}$ are the mean of the observed and forecasted values, respectively. $n$ is the total number of time steps. $H$, $M$ and $F$ are the number of hits (correctly predicted events), misses (events that occurred but were not predicted) and false alarms (events that were predicted but did not occur). The 95th percentiles of multi-year daily precipitation forecasts and observations at each forecast lead time are used as thresholds to distinguish extreme precipitation events in the forecasts and observations respectively.

TCC primarily assesses the linear relationship between the observation and the forecast, without considering absolute value discrepancies. Its range is $[-1, 1]$, with values approaching 1 indicating a stronger linear association between the two. RMSE quantifies the discrepancy between the observation and the forecast, with values ranging from $[0, +\infty)$ , where values closer to 0 signify smaller errors. CSI is a metric used to evaluate the accuracy in detecting binary events (extreme precipitation events in this study), with a range of $[0,1]$, where values closer to 1 indicate a greater ability to accurately predict extreme precipitation events (Jolliffe and Stephenson, 2011). Unlike the widely used metrics probability of detection (POD) and false alarm ratio (FAR), the CSI takes into account both false alarms and missed events, providing a more balanced assessment metric.

### 3.3.3. Ensemble forecast evaluation metrics

For the performance evaluation of ensemble forecasts (multi-point), following the studies by Y. Zhang et al. (2023) and Klotz et al.(2022), we assess from three perspectives: overall performance, reliability, and sharpness. Overall performance should encompass both reliability and sharpness, reflecting the ensemble forecast's overall capability in representing observations comprehensively. Reliability measures how consistent the provided uncertainty estimates are with respect to the available observations. A higher model resolution does not necessarily ensure greater reliability as the model should strike a balance between precision and accuracy, avoiding both excessive confidence (over-confidence) and excessive dispersion (under-confidence). Sharpness describes the precision or concentration of a probabilistic prediction, indicating how well the predicted probability distributions correspond to the observations. A sharper forecast suggests narrower predicted uncertainties that closely match the observed data, providing a more accurate depiction of the true uncertainty in the predictions (Klotz et al., 2022).

Continuous ranked probability score (CRPS) is chosen as the metric to evaluate overall performance because it simultaneously considers the calibration of the predictive distribution (whether the predictive distribution covers the true values) and the sharpness (Gneiting, 2008). The calculation for CRPS is shown in formula (14). CRPS calculates the degree of match between the entire predictive distribution and the true value, with a range of $[0, +\infty)$. A smaller CRPS indicates that the predictive distribution is closer to the true value and that uncertainty is controlled reasonably. If the predictive distribution is overly dispersed, such as having a large variance, or if the mean deviates significantly from the true value, the CRPS will increase. The CRPS at each grid point is defined as the average CRPS computed over all time steps at that specific grid location.

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - 1(x \geq y))^2 dx \tag{16}$$

Where $F(x)$ is the CDF of the predictive distribution, $y$ is the true observation, and $1(x \geq y)$ is the indicator function for the true value (which is 1 when $x \geq y$, and 0 otherwise).

We employ the Reliability Diagram (also referred to as Probability Plots) to assess the reliability between predicted probabilities and observed frequencies. In this diagram, the theoretical quantiles of a uniform distribution are plotted along the x-axis, while the proportion of observed values falling below the corresponding predictions is plotted on the y-axis. Ideally, in a perfectly reliable forecast, if an event is predicted with a probability of 50 %, its observed relative frequency should also approximate 50 %. Consequently, data points in the reliability diagram of a perfect forecast should align along the diagonal, indicating a consistent match between predicted probabilities and observed frequencies across different probability levels. Deviations from this 1:1 diagonal highlight potential bias in the model. If points lie above the diagonal, the observed relative frequency exceeds the predicted probability, suggesting an underprediction phenomenon. Conversely, if points fall below the diagonal, the observed relative frequency is lower than the predicted probability, indicating an overprediction tendency. For further details regarding the principles and application of reliability diagrams, refer to Jolliffe and Stephenson (2011) and Laio and Tamea

(2007).

The 50 % and 95 % percentile intervals are selected as the evaluation criteria for sharpness. Specifically, we measured the average Euclidean distance of all time steps between the 25 % and 75 % quantiles ($DIS_{50}$) and between the 2.5 % and 97.5 % quantiles ($DIS_{95}$) within the ensemble members. Additionally, we calculated the ratio of observed values falling within the corresponding prediction intervals to the total number of observations, denoted as $CO_{50}$ and $CO_{95}$. To enhance the assessment of sharpness across the full ensemble, we further computed three commonly used metrics: mean absolute deviation (MAD), standard deviation (SD), and variance (VAR), based on all ensemble members. For a sharper forecast, we expect lower values for $DIS_{50}$, $DIS_{95}$, MAD, SD, and VAR, indicating reduced dispersion and tighter predictive intervals. Meanwhile, higher values for $CO_{50}$ and $CO_{95}$ are desirable, reflecting better coverage of observations within the corresponding prediction intervals. These metrics are finally averaged over all grid points.

### 3.4. Interpretability analysis method: SHAP analysis

Machine learning methods have long been criticized as "black boxes" due to their complexity and lack of interpretability. To further enhance the interpretability of the PDML model, we choose to use a Post-hoc techniques Shapley Additive Explanations (SHAP) analysis method to explain the contribution of each selected feature to the output parameters (Tripathy and Mishra, 2024). SHAP analysis is based on Shapley values from game theory, which are used to measure the contribution of each feature to the model's prediction outcome (Lundberg and Lee, 2017). SHAP analysis helps us understand how individual features influence the model's predictions and provides a quantitative explanation of the relative importance of different features. By calculating the contribution of each feature to the model's output, SHAP produces a SHAP value. The SHAP value can be positive or negative, indicating the direction of the feature's influence on the prediction (positive values increase the prediction, while negative values decrease it). The formula for calculating the Shapley value is given by equation (15). This analytical approach was not employed in previous DML studies because DML directly overlays the classifier results with the regressor results, making its inherent logic and generated outcomes difficult to interpret.

$$\phi_i = \sum_{S \in N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f(S \cup \{i\}) - f(S)] \tag{17}$$

where $N$ is the set of features $\{1, 2, \cdots, M\}$. $S$ is a subset of the feature set $N$ that does not include feature $i$, and $|S|$ is the size of subset $S$ (i.e., the number of features in the subset). $f(S)$ is the model's output when only using the feature set $S$, and $f(S \cup \{i\})$ is the model's output when using the feature set $S$ and feature $i$.

Considering that the focus of this paper is not on explainable machine learning, the SHAP analysis is conducted solely to facilitate understanding of how features influence the output parameters in PDML models. Therefore, we use the PDML-GRU model with a 1-day lead time as an example and perform the analysis on only two grid points.

## 4. Results and discussion

### 4.1. Classification accuracy validation

The classification ACC results of the five models are presented in Table 1. As shown, the RNN-based LSTM and GRU models consistently achieve higher accuracy across all lead times, while U-NET and KAN exhibit slightly inferior classification performance compared to RNN architectures. This also indirectly demonstrates that RNN-based models are better suited for classification tasks. Overall, all models except MLP demonstrate satisfactory classification accuracy—exceeding 0.7 for daily precipitation forecasts and surpassing 0.8 for accumulated precipitation forecasts—a pattern that aligns with their subsequent dominance over MLP in comprehensive validation tests. As previously introduced, however, ACC should be interpreted as a only supplementary metric only consider the class with maximum predicted probability. While class imbalance in precipitation events inevitably leads DL models to exhibit higher probability outputs for majority classes, this does not imply that minority-class probabilities are neglected in PDML. Rather, PDML systematically incorporates these probabilities into the regression framework through weighted integration, where critical minority events are assigned elevated weights to preserve their physical significance in the final prediction system. Thus, classification accuracy must be evaluated in conjunction with regressor performance. For instance, despite its lower classification ACC relative to RNN-based models, U-NET emerges as the optimal model in accumulated-scale validation studies, underscoring the compensatory mechanisms enabled by its robust probabilistic outputs and regression components.

**Table 1**
Classification Accuracy of daily and accumulated precipitation forecasts.

| | | LSTM | GRU | UNET | MLP | KAN |
|---|---|---|---|---|---|---|
| Daily precipitation forecasts | Lead day 1 | 0.76 | 0.76 | 0.72 | 0.60 | 0.73 |
| | Lead day 2 | 0.76 | 0.76 | 0.72 | 0.60 | 0.73 |
| | Lead day 3 | 0.76 | 0.76 | 0.72 | 0.59 | 0.73 |
| | Lead day 4 | 0.75 | 0.76 | 0.71 | 0.59 | 0.73 |
| Accumulated precipitation forecasts | Lead days 1–7 | 0.82 | 0.82 | 0.80 | 0.61 | 0.82 |
| | Lead days 8–14 | 0.81 | 0.81 | 0.79 | 0.50 | 0.80 |
| | Lead days 15–30 | 0.86 | 0.86 | 0.86 | 0.48 | 0.85 |
| | Lead days 31–60 | 0.88 | 0.88 | 0.86 | 0.54 | 0.87 |

The evaluation results of CE Loss are presented in **Table. S2** of the supplementary materials. These results are generally consistent with those of ACC, with the RNN-based model performing the best, followed by U-NET and KAN, while the MLP model exhibits a performance deviation.

### 4.2. Deterministic assessment

The evaluation results of average TCC, RMSE (mm/d) and CSI across all grids for the raw forecast and different post-processing methods are shown in Table 2. The best-performing metrics within the category are highlighted in bold. The results indicate that for all forecast periods, EPP and the five PMDL post-processing methods achieve significant improvements over the raw forecast, as evidenced by higher TCC, CSI and lower RMSE. However, there are differences in performance across the various machine learning algorithms within the PDML framework. While it is evident that the top-performing three models consistently emerge from LSTM, GRU, and U-NET, the best-performing post-processing algorithm varies across different time scales.

For the TCC and RMSE of daily precipitation forecast, we found that for forecast periods of 1, 2, and 3 day, the RNN-based machine learning algorithms performed the best within the PMDL framework, with GRU slightly outperforming LSTM. Compared to the raw forecast, the correlation coefficient increased by up to 68.7 %, 69.6 %, and 71.1 %, respectively, while RMSE decreased by up to 44.7 %, 44.6 %, and 43.3 %. However, for the 4-day lead, the U-NET algorithm performed the best within the PDML framework, with the correlation coefficient improving by 76.0 % and RMSE decreasing by 41.2 %. For the accumulated precipitation forecast, we found that for the forecast periods of 1–7 days, 8–14 days, and 31–60 days, the U-NET-based PDML performed the best, with correlation coefficients improving by 39.7 %, 126.7 %, and 102.1 %, respectively, and RMSE reducing by 46.6 %, 44.9 %, and 49.3 %. LSTM performs best in the lead days 15–30, with a 132.3 % increase in TCC and a 48.4 % decrease in RMSE. RNN-based and U-NET algorithm outperformed the EPP method across all forecast periods, while other algorithms outperformed EPP only in specific forecast periods or performed worse than EPP across all forecast periods. Particularly, the MLP

**Table 2**
Results of deterministic evaluation metrics.

| | Lead | Metrics | RAW | EPP | LSTM | GRU | UNET | MLP | KAN |
|---|---|---|---|---|---|---|---|---|---|
| Daily precipitation forecasts | Lead day 1 | TCC | 0.37 | 0.51 | 0.62 | **0.63** | 0.57 | 0.50 | 0.53 |
| | | RMSE | 5.81 | 3.56 | 3.26 | **3.22** | 3.63 | 4.70 | 3.53 |
| | | CSI | 0.07 | 0.19 | 0.27 | **0.28** | 0.24 | 0.18 | 0.21 |
| | Lead day 2 | TCC | 0.37 | 0.50 | 0.62 | **0.63** | 0.56 | 0.49 | 0.53 |
| | | RMSE | 5.83 | 3.56 | 3.28 | **3.24** | 3.65 | 4.74 | 3.54 |
| | | CSI | 0.07 | 0.19 | **0.27** | **0.27** | 0.22 | 0.18 | 0.21 |
| | Lead day 3 | TCC | 0.36 | 0.50 | **0.61** | **0.61** | 0.57 | 0.48 | 0.52 |
| | | RMSE | 5.82 | 3.58 | 3.32 | **3.3** | 3.65 | 4.82 | 3.56 |
| | | CSI | 0.07 | 0.18 | **0.26** | **0.26** | 0.23 | 0.17 | 0.20 |
| | Lead day 4 | TCC | 0.33 | 0.48 | **0.57** | **0.57** | 0.58 | 0.47 | 0.49 |
| | | RMSE | 5.84 | 3.63 | **3.44** | 3.47 | 3.46 | 4.66 | 3.65 |
| | | CSI | 0.07 | 0.18 | 0.23 | 0.23 | **0.24** | 0.16 | 0.19 |
| Accumulated precipitation forecasts | Lead days 1–7 | TCC | 0.54 | 0.70 | 0.73 | 0.71 | **0.75** | 0.70 | 0.70 |
| | | RMSE | 2.91 | 1.64 | 1.57 | 1.69 | **1.56** | 1.83 | 1.72 |
| | | CSI | 0.09 | 0.24 | 0.26 | 0.26 | **0.28** | 0.22 | 0.26 |
| | Lead days 8–14 | TCC | 0.28 | 0.60 | 0.62 | 0.61 | **0.63** | 0.59 | 0.59 |
| | | RMSE | 3.30 | 1.85 | **1.82** | 1.87 | **1.82** | 2.15 | 1.92 |
| | | CSI | 0.04 | 0.13 | 0.15 | 0.16 | **0.16** | 0.14 | 0.14 |
| | Lead days 15–30 | TCC | 0.31 | 0.68 | **0.72** | 0.70 | 0.71 | 0.68 | 0.67 |
| | | RMSE | 2.50 | 1.36 | **1.29** | 1.38 | 1.31 | 1.56 | 1.48 |
| | | CSI | 0.03 | 0.11 | **0.16** | 0.16 | 0.14 | 0.14 | 0.13 |
| | Lead days 31–60 | TCC | 0.39 | 0.75 | 0.79 | 0.76 | **0.80** | 0.76 | 0.72 |
| | | RMSE | 1.97 | 1.07 | **0.99** | 1.08 | 1.00 | 1.16 | 1.25 |
| | | CSI | 0.04 | 0.12 | **0.17** | 0.16 | 0.16 | 0.14 | 0.14 |

model demonstrates significantly lower performance within the PDML framework compared to other models, although it still manages to improve the raw forecasts.

For the CSI of daily precipitation forecast, the RNN-based PDML methods perform best on lead day 1, 2, and 3, with GRU slightly outperforming LSTM, achieving maximum CSI improvements of 284.5 %, 282.4 %, and 276.6 %, respectively. On lead day 4, U-NET performs best, improving CSI by 269.8 %. Across all lead times, LSTM, GRU, U-NET, and KAN consistently demonstrate better performance compared to EPP. For the accumulated precipitation forecast, U-NET performs best on lead days 1–7 and 8–14, while GRU and LSTM achieve the best performance on lead days 15–30 and 31–60, respectively. The four lead periods improve CSI by 208.5 %, 295.3 %, 393.0 %, and 346.7 %, respectively. Except for MLP, the other four PDML methods outperform EPP across all lead periods, while MLP also surpasses EPP during lead days 8–60. The significant improvement in CSI proves that PDML can effectively enhance the ability to predict the occurrence of extreme precipitation events.

In particular, we also explored the spatial features of improvement made by each model. The spatial distribution of TCC improvements for daily precipitation forecasts and accumulated precipitation forecasts are shown in Fig. 3 and Fig. 4, respectively. At the daily scale, we found that the improvements of PDML compared to EPP were most significant in the central and northwest regions, but relatively small in the South China and Tibetan Plateau regions. In South China, the daily precipitation characteristics are complex, and the deep learning models were unable to fully capture the precipitation mapping features. In contrast, for the Tibetan Plateau, the original forecasts performed relatively accuracy, and post-processing models showed limited improvement (as shown in Fig. S2-Fig. S3 in the supplementary materials). At the accumulated scale, the pattern differed. PDML showed the most significant improvements in the Tibetan Plateau and North China regions, with only the U-NET model demonstrating noticeable improvements in the central and South China regions. This may be due to the fact that for accumulated precipitation, the sliding average over each time step reduces internal differences, offering limited information for RNN-based models, whereas the U-NET model is better at capturing spatial patterns. The South China region is the area most affected by the East Asian Summer Monsoon, especially for longer-scale subseasonal precipitation, where the spatial pattern of the monsoon's northward influence is significant (Wang et al., 2009). The U-NET model may improve the accumulated precipitation forecasts in South China by capturing the spatial patterns of multi-scale monsoon dynamics. Similarly, the improvement in RMSE also exhibits similar spatial distribution patterns. Interested readers can refer to Fig. S4-Fig. S5 in the supplementary materials.

The same improvement pattern also appears in the CSI metric (see Fig. S6-Fig. S7 in the supplementary materials), with more noticeable improvements at the daily scale. Of course, our primary focus is on extreme precipitation at the daily scale, as the moving average over longer time scales tends to dilute the short-term concentration of extreme precipitation. Improvements at the daily scale are primarily concentrated in the central, eastern, northwest, and northeastern regions, while the improvements in the South China region are relatively insufficient. On the accumulated scale, the improvements are mainly concentrated in the Qinghai-Tibet Plateau, North China, and the Northeast region. We believe the limited improvement in accumulated precipitation forecasts is due to the fact that after averaging over longer time scales, the extreme characteristics of precipitation (such as short-duration, high-intensity events) become less pronounced. When placed on a longer time scale, these extreme features are diluted, making it difficult for PDML to effectively distinguish between extreme precipitation and general precipitation events.

To compare the seasonal differences in the post-processing performance of various models, the evaluation results of TCC and RMSE for four different seasons (March-May, MAM; June-August, JJA; September-November, SON; December-February, DJF) are presented in Table 3. We found that the RNN-based and U-NET models in PDML outperformed EPP in all seasons, with a significant improvement compared to the original forecast. For TCC, we observed that the improvement was greater in the autumn and winter seasons (SON and DJF) than in the spring and summer seasons (MAM and JJA), possibly because mainland China begins to experience the influence of monsoons during the spring and summer, leading to higher variability and uncertainty in precipitation within the season. For RMSE, the improvement was more pronounced in winter than in summer. Considering that the RMSE of the original forecast was nearly the same across all seasons, we believe this improvement was mainly due to the removal of systematic biases within the season. The relatively simpler MLP model clearly lacked sufficient capability and even introduced larger errors in the summer.

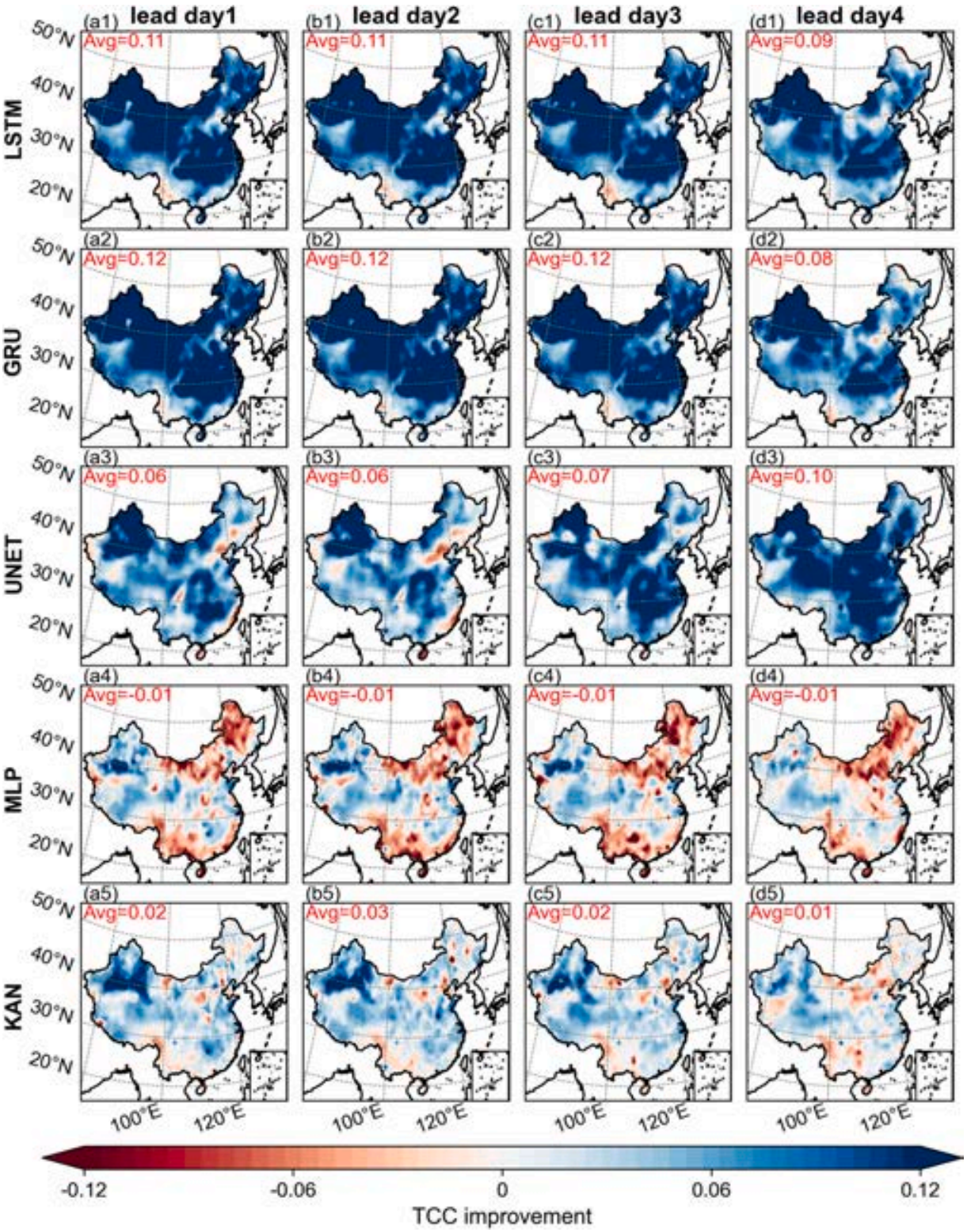



**Fig. 3.** Spatial distribution of the difference in TCC between PDML models and EPP of daily precipitation forecasts, with blue representing improvement. (a1)-(a5) lead day1; (b1)-(b5) lead day2; (c1)-(c5) lead day3; (d1)-(d5) lead day4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### *4.3. Probabilistic assessment*

The evaluation results of CRPS for each forecast period are shown in Table 4. The best-performing CRPS for each lead time are highlighted in bold. The average CRPS is obtained by averaging the CRPS values across all grid points, while the CRPS at each grid point is computed as the average over all time steps. As the original forecast utilized in this study is a control forecast, which represents a deterministic single-point prediction, the computation of CRPS is not applicable. Consequently, the EPP method remains the only viable benchmark for performance

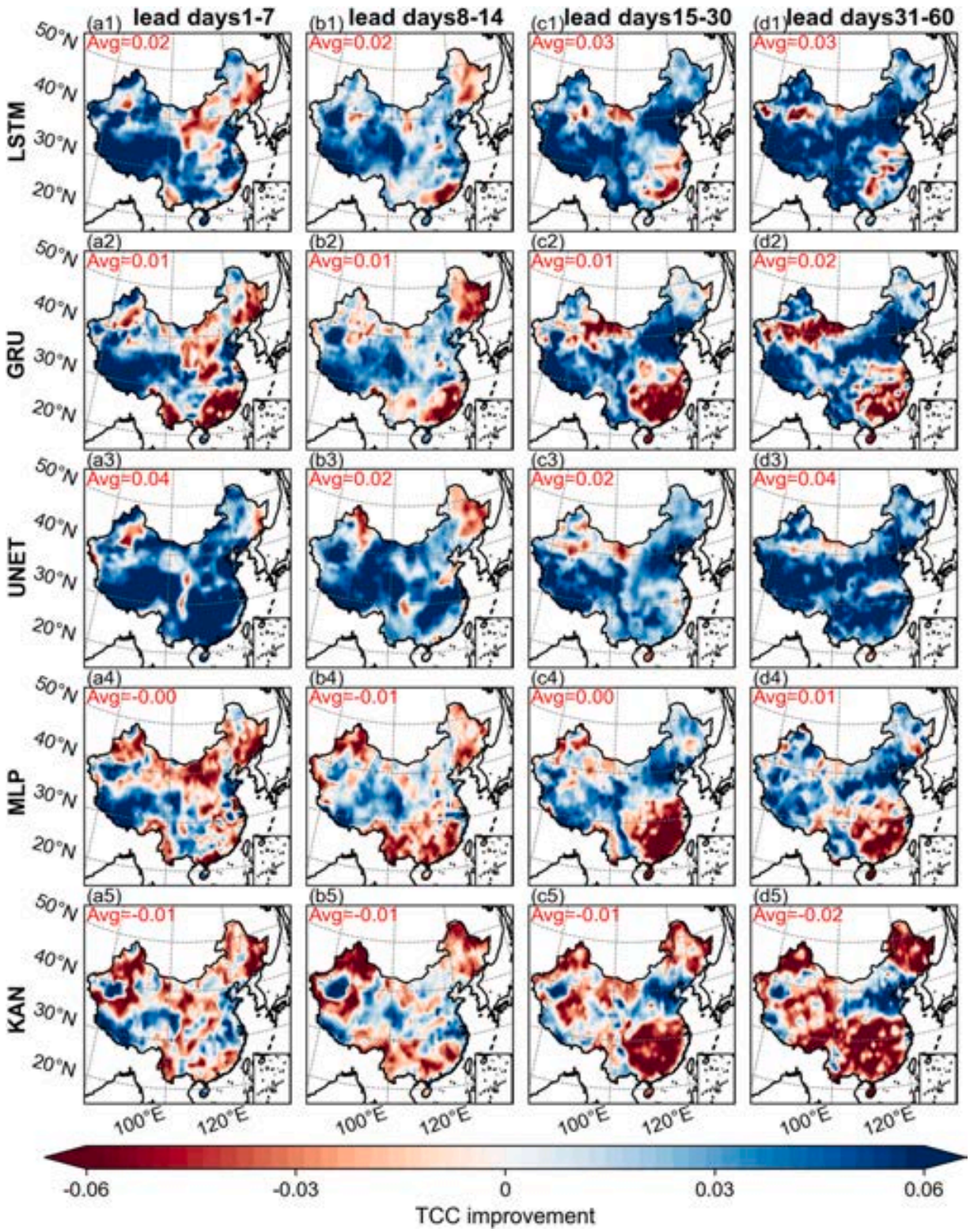



**Fig. 4.** Spatial distribution of the difference in TCC between PDML models and EPP of accumulated precipitation forecasts, with blue representing improvement. (a1)-(a5) lead days1-7; (b1)-(b5) lead days8-14; (c1)-(c5) lead days15-30; (d1)-(d5) lead days31-60. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

evaluation. For daily forecasts, both RNN-based and U-NET structured PDML methods achieve superior performance compared to the EPP method. Specifically, at lead times of day 1, 2, 3, and 4, the CRPS can be reduced by up to 13.4 %, 14.2 %, 13.0 %, and 8.4 %, respectively, relative to EPP. For accumulated forecasts, both U-NET and LSTM show robust improvements compared to EPP, with a reduction in CRPS of 4.2 %, 1.4 %, 6.0 % and 9.7 % across four lead times. MLP and KAN both underperform relative to EPP at accumulated scale while only KAN is better than EPP in daily scale. We found that although the CRPS performance is better with the RNN-based and U-NET models, the ranking

**Table 3**
Results of deterministic and probabilistic evaluation metrics for different seasons on lead day 1 and lead days 1–7.

| | Season | Metrics | RAW | EPP | LSTM | GRU | UNET | MLP | KAN |
|---|---|---|---|---|---|---|---|---|---|
| Lead day 1 precipitation forecasts | MAM | TCC | 0.37 | 0.43 | 0.59 | **0.61** | 0.57 | 0.44 | 0.45 |
| | | RMSE | 5.81 | 2.91 | 2.58 | **2.54** | 2.65 | 3.60 | 2.88 |
| | | CRPS | Nan | 1.05 | 0.88 | **0.87** | 0.90 | 1.33 | 1.02 |
| | JJA | TCC | 0.37 | 0.39 | 0.52 | **0.54** | 0.53 | 0.38 | 0.44 |
| | | RMSE | 5.83 | 5.27 | 4.94 | **4.88** | 4.98 | 6.84 | 5.23 |
| | | CRPS | Nan | 2.12 | 1.91 | 1.89 | **1.85** | 2.89 | 2.04 |
| | SON | TCC | 0.36 | 0.45 | **0.60** | **0.60** | 0.57 | 0.46 | 0.48 |
| | | RMSE | 5.82 | 3.13 | 2.79 | **2.75** | 2.90 | 4.13 | 3.07 |
| | | CRPS | Nan | 1.04 | 0.90 | **0.89** | **0.89** | 1.45 | 1.01 |
| | DJF | TCC | 0.33 | 0.33 | 0.52 | **0.54** | 0.50 | 0.34 | 0.36 |
| | | RMSE | 5.84 | 1.36 | 1.22 | **1.21** | 1.27 | 2.16 | 1.40 |
| | | CRPS | Nan | 0.38 | 0.33 | **0.32** | 0.33 | 0.61 | 0.40 |
| Lead days 1–7 precipitation forecasts | MAM | TCC | 0.48 | 0.59 | 0.60 | 0.57 | **0.65** | 0.56 | 0.53 |
| | | RMSE | 2.52 | 1.27 | 1.25 | 1.43 | **1.24** | 1.46 | 1.37 |
| | | CRPS | Nan | 0.63 | 0.63 | 0.71 | **0.62** | 0.74 | 0.72 |
| | JJA | TCC | 0.46 | 0.49 | 0.54 | 0.53 | **0.59** | 0.49 | 0.53 |
| | | RMSE | 4.28 | 2.52 | 2.40 | 2.49 | **2.37** | 2.73 | 2.60 |
| | | CRPS | Nan | 1.31 | 1.22 | **1.20** | 1.22 | 1.48 | 1.40 |
| | SON | TCC | 0.51 | 0.62 | 0.65 | 0.63 | **0.69** | 0.62 | 0.59 |
| | | RMSE | 2.45 | 1.36 | **1.30** | 1.42 | **1.30** | 1.61 | 1.42 |
| | | CRPS | Nan | 0.63 | **0.60** | 0.62 | **0.60** | 0.74 | 0.69 |
| | DJF | TCC | 0.46 | 0.48 | 0.47 | 0.45 | **0.55** | 0.42 | 0.41 |
| | | RMSE | 1.19 | 0.60 | 0.60 | 0.66 | **0.58** | 0.74 | 0.84 |
| | | CRPS | Nan | 0.26 | 0.26 | 0.27 | **0.25** | 0.31 | 0.35 |

**Table 4**
Results of overall ensemble evaluation metrics CRPS.

| | Lead | EPP | LSTM | GRU | UNET | MLP | KAN |
|---|---|---|---|---|---|---|---|
| Daily precipitation forecasts | Lead day 1 | 1.15 | 1.01 | **1.00** | **1.00** | 1.58 | 1.12 |
| | Lead day 2 | 1.15 | 1.01 | 1.00 | **0.98** | 1.58 | 1.12 |
| | Lead day 3 | 1.15 | 1.03 | 1.02 | **1.00** | 1.60 | 1.13 |
| | Lead day 4 | 1.17 | **1.07** | 1.11 | 1.11 | 1.56 | 1.19 |
| Accumulated precipitation forecasts | Lead days 1–7 | 0.71 | **0.68** | 0.71 | **0.68** | 0.82 | 0.79 |
| | Lead days 8–14 | 0.83 | **0.81** | 0.83 | **0.81** | 1.00 | 0.88 |
| | Lead days 15–30 | 0.66 | **0.62** | 0.67 | 0.63 | 0.80 | 0.74 |
| | Lead days 31–60 | 0.55 | **0.49** | 0.56 | **0.49** | 0.63 | 0.66 |

**Table 5**
Sharpness statistics of daily and accumulated precipitation forecasts.

| | | EPP | LSTM | GRU | UNET | MLP | KAN |
|---|---|---|---|---|---|---|---|
| Daily precipitation forecasts | $DIS_{50}$ | 1.91 | 1.81 | 1.77 | 1.19 | 3.64 | 1.92 |
| | $DIS_{95}$ | 8.32 | 7.72 | 7.42 | 4.11 | 12.0 | 8.40 |
| | $CO_{50}$ | 0.71 | 0.67 | 0.67 | 0.57 | 0.66 | 0.68 |
| | $CO_{95}$ | 0.96 | 0.96 | 0.96 | 0.82 | 0.97 | 0.96 |
| | MAD | 0.70 | 0.66 | 0.65 | 0.50 | 1.30 | 0.64 |
| | SD | 2.41 | 2.28 | 2.19 | 1.15 | 3.59 | 2.45 |
| | VAR | 15.21 | 14.27 | 13.46 | 4.83 | 31.40 | 15.62 |
| Accumulated precipitation forecasts | $DIS_{50}$ | 1.57 | 1.27 | 1.15 | 1.23 | 2.05 | 1.99 |
| | $DIS_{95}$ | 5.06 | 4.06 | 3.88 | 3.72 | 6.35 | 5.84 |
| | $CO_{50}$ | 0.62 | 0.55 | 0.55 | 0.55 | 0.62 | 0.66 |
| | $CO_{95}$ | 0.97 | 0.94 | 0.92 | 0.93 | 0.98 | 0.98 |
| | MAD | 0.70 | 0.55 | 0.48 | 0.55 | 0.83 | 0.79 |
| | SD | 1.37 | 1.12 | 1.14 | 1.02 | 1.82 | 1.67 |
| | VAR | 4.15 | 2.72 | 2.08 | 2.42 | 6.66 | 4.82 |

does not align perfectly with that of the deterministic metrics. This phenomenon has also been observed in similar probabilistic forecasting studies. The reason is that CRPS not only considers the difference in means but also the distance between the entire probability distribution and the true values (Jahangir and Quilty, 2024). Excessive ensemble spread can lead to a decline in CRPS, and the ensemble spread metric is analyzed in the subsequent Sharpness section (Table 5).

The spatial distribution of the CRPS difference between PDML and EPP is shown in Fig. 5 and Fig. 6. We find that at the daily scale, both the RNN-based and U-NET models show improvements over EPP in almost all regions of mainland China, with the most significant improvements observed in the central and eastern regions. The MLP model fails to show any improvement in any region, while KAN demonstrates improvements in most areas, except for the southwest. However, the improvement is much smaller compared to the more sophisticated PDML models. At the accumulated scale, the improvements from the PDML models are relatively smaller, with the improvements primarily driven by the LSTM and U-NET models. The improvements from both models are mainly distributed in the Qinghai-Tibet Plateau and North China regions, while U-NET shows a noticeable improvement in the central region, and LSTM shows a significant improvement in the southwest. In South China, both the RNN-based and U-NET models lack significant improvements, especially the RNN-based model, which even shows a large area of performance degradation in the coastal regions of the south. We believe this may be related to the previously mentioned monsoon activity, as the RNN-based model is not adept at capturing large-scale spatial patterns, and the additional information provided by accumulated precipitation forecasts over time steps is limited.

The CRPS for each model in different seasons is also shown in Table 3. We found that, compared to EPP, PDML models, except for MLP, generally lead to improvements at the daily scale in all seasons, which demonstrates the broad reliability of PDML models in generating probabilistic forecasts. However, at the accumulated scale, only the CRPS for summer shows significant improvement. Regarding this phenomenon, we believe it is primarily due to the fact that the calculation of CRPS depends on the size of the numerical probability distribution range. For accumulated precipitation in seasons other than summer, after multi-day averaging, the numerical base is relatively small, and the ensemble range is not as wide as in summer, making it difficult to observe significant improvements.

To assess the reliability of ensemble forecasts for different levels of precipitation, we set three different thresholds to plot the reliability

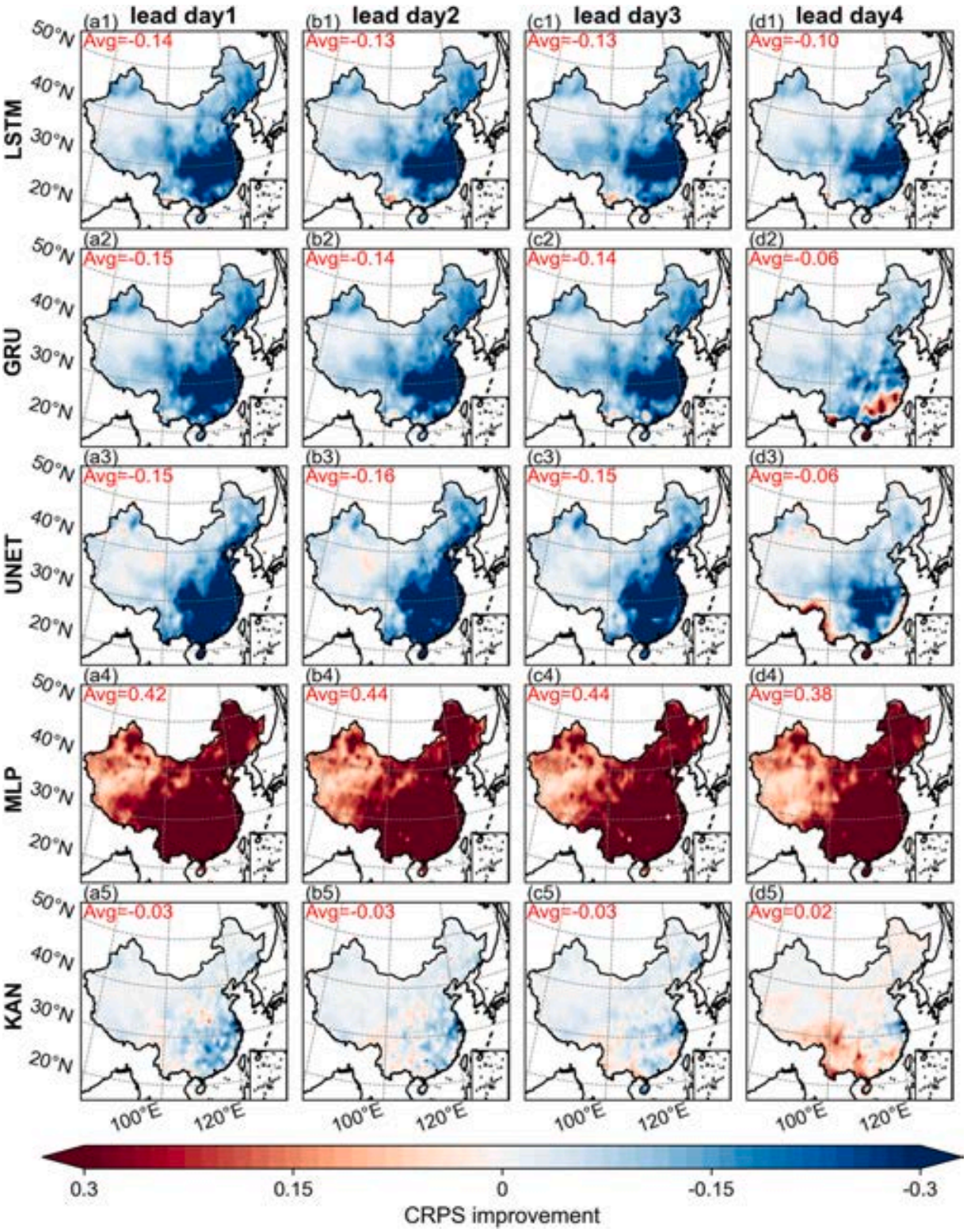



**Fig. 5.** Spatial distribution of the difference in CRPS between PDML models and EPP of daily precipitation forecasts, with blue representing improvement. (a1)-(a5) lead day1; (b1)-(b5) lead day2; (c1)-(c5) lead day3; (d1)-(d5) lead day4. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

diagram. Considering the significant variability in precipitation across mainland China, we flattened the observation-ensemble forecast pairs for all grid points into a one-dimensional array. Then, we selected the 50th, 80th, and 95th percentiles of the observations for reliability assessment. The results of the reliability diagram are shown in Fig. 7. We found that for daily precipitation forecasts, when the threshold is set at the 50th percentile, the EPP method shows some overprediction at lead day 4, while the U-NET method exhibits underprediction in the low-frequency region for lead times 2, 3, and 4 days, and overprediction in the high-frequency region. The performance of the other methods is

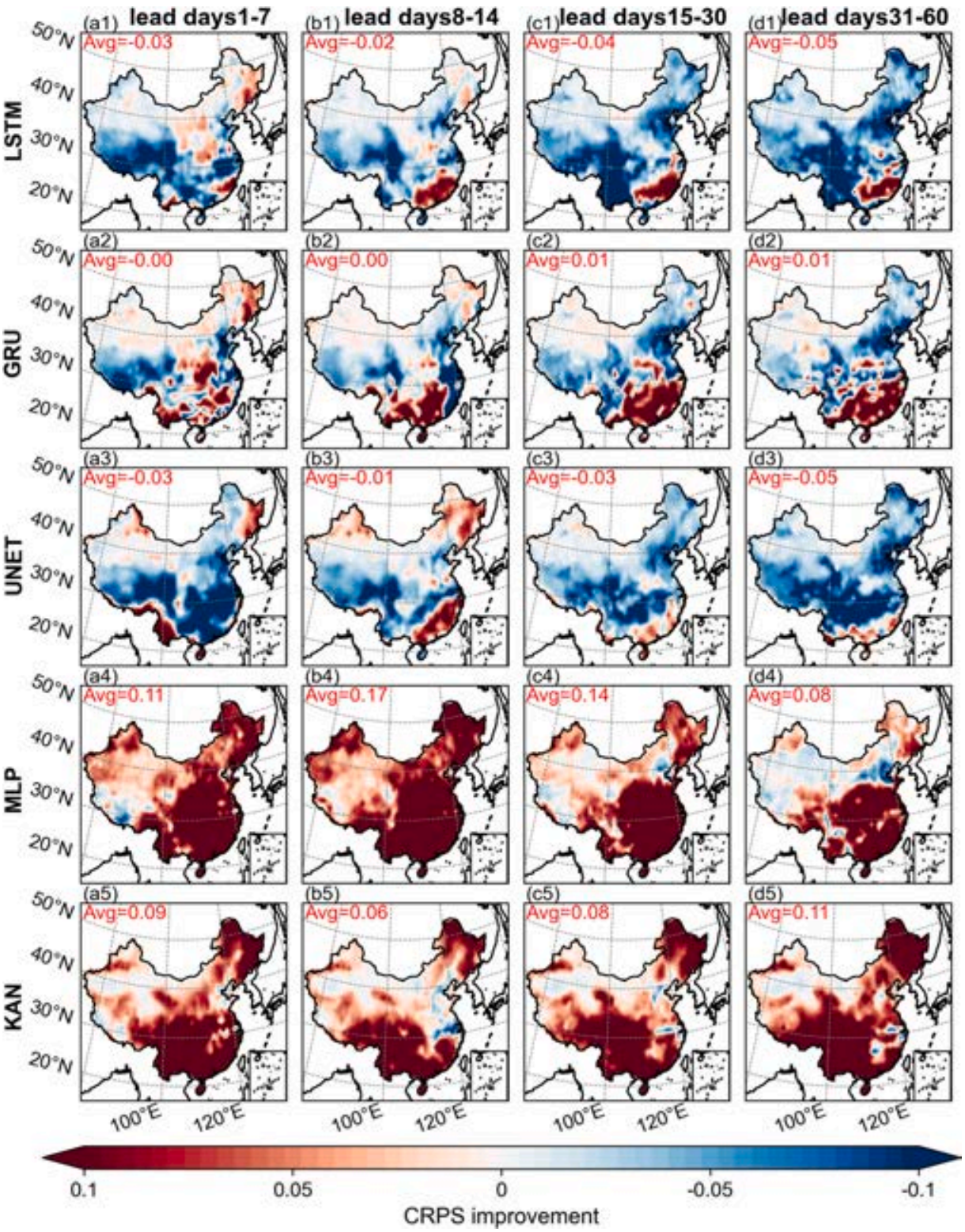



**Fig. 6.** Spatial distribution of the difference in CRPS between PDML models and EPP of accumulated precipitation forecasts, with blue representing improvement. (a1)-(a5) lead days1-7; (b1)-(b5) lead days8-14; (c1)-(c5) lead days15-30; (d1)-(d5) lead days31-60. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

generally reliable, clustering around the 1:1 line. When the threshold is set at the 80th percentile, MLP consistently shows overprediction, while the other methods perform relatively reliably. At the 95th percentile threshold, MLP still exhibits significant overprediction, while U-NET shows underprediction at lead times 2, 3, and 4 days. For accumulated precipitation forecasts, we found that various methods exhibit overall more stable performance compared to daily precipitation forecasts. Except for the LSTM and KAN methods, which show noticeable underprediction at the 80th percentile threshold, the other methods remain close to the 1:1 line in all other cases.

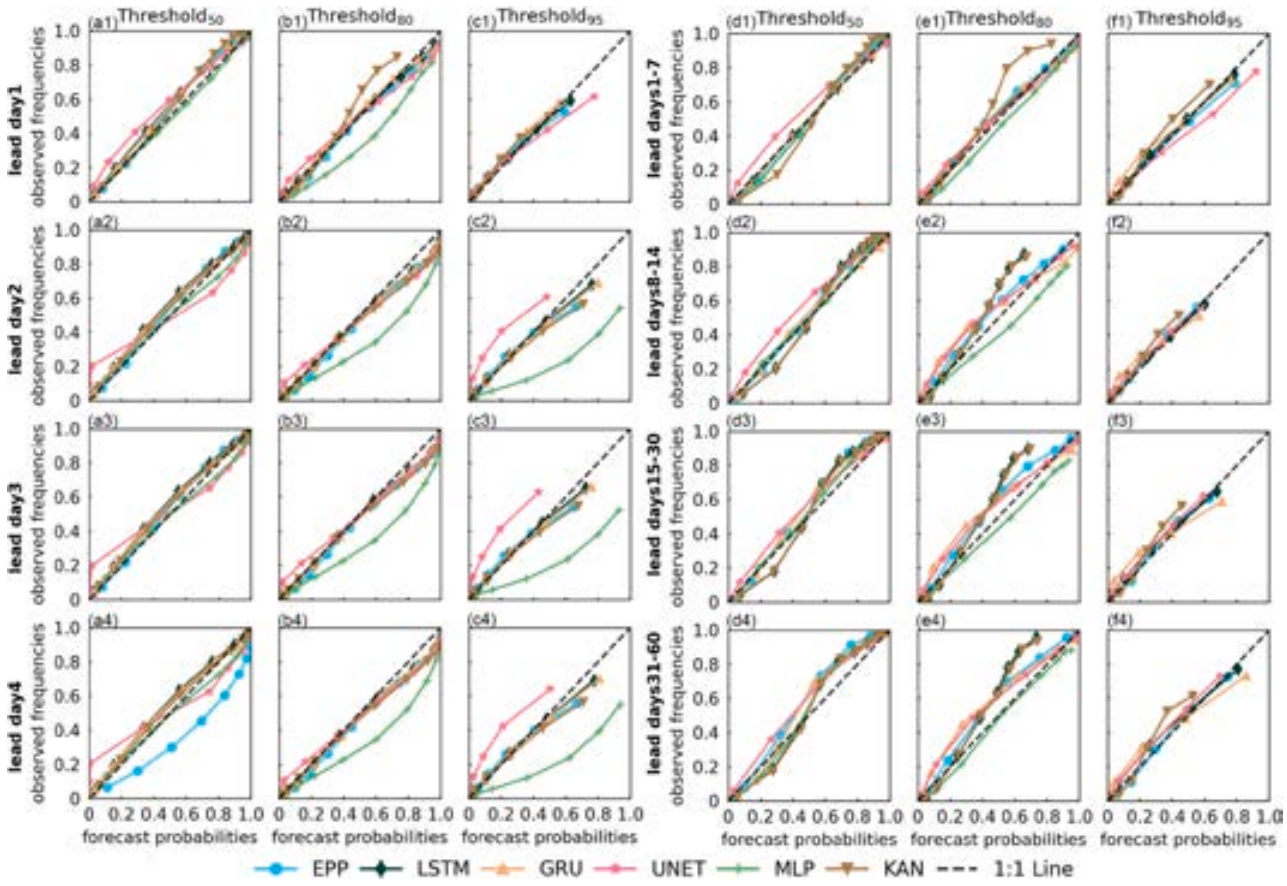



**Fig. 7.** Reliability diagrams of various post-processed ensemble forecasts at three different thresholds (50th percentile, 80th percentile, and 95th percentile). (a1-c1), (a2-c2), (a3-c3), and (a4-c4) represent the reliability diagrams of four daily precipitation forecasts at the three thresholds; (d1-f1), (d2-f2), (d3-f3), and (d4-f4) represent the reliability diagrams of four accumulated precipitation forecasts at the three thresholds. The results in this figure are calculated by flattening data of all grids into one-dimensional array.

The reliability diagram reveals some deficiencies in the various PDML methods. For instance, although the U-NET method shows significant advantages in accumulated precipitation forecasts, it exhibits a clear inability to capture high precipitation values (95th percentile) in daily precipitation forecasts. This may be due to the limitations of using a small sample size, where the spatial convolution in the U-NET model leads to the model capturing precipitation forecast information at a larger scale, which results in insufficient ability to capture high precipitation values at finer scales. At the same time, the MLP method shows a clear overprediction in daily precipitation forecasts. This is likely due to the simplicity of the model, combined with the fact that we did not apply a loss function specifically targeting medium-to-high precipitation values. This leads to a common issue in machine learning models, where they tend to perform well at capturing precipitation near the mean but lack the ability to map high-value precipitation effectively (Larraondo et al., 2020).

The statistical metrics for sharpness are shown in Table 4. We calculated the averages for the selected daily precipitation forecasts and accumulated precipitation forecasts across each four lead times, respectively. For daily precipitation forecasts, it was found that the U-NET architecture in the PDML method has the smallest ensemble spread, as evidenced by the lowest values of DIS, MAD, SD, and VAR. At the same time, the smaller ensemble spread also makes the U-NET PDML model struggle to capture a sufficient number of precipitation events. On the other hand, both LSTM and GRU models, while covering a similar number of precipitation events as EPP, exhibit smaller ensemble spreads, thus achieving a sharper result. Although MLP and KAN also manage to cover a sufficient number of precipitation events, they show significantly larger ensemble spreads, leading to excessive uncertainty in the generated results, which is less beneficial for decision-makers trying to gather actionable information. For accumulated precipitation forecasts, both the RNN-based and U-NET models show a significant reduction in ensemble spread compared to EPP (over a 20 % decrease), but the proportion of observed precipitation they cover is only slightly lower than EPP, by approximately 10 %. This also reflects the improvement in the general performance metrics CRPS.

## 5. Interpretability and feature contributions

The two grid points randomly selected in this study are 117°E 29°N and 84°E 39°N, with the former located in the humid region of southeastern China and the latter in the arid region of the northwest. The SHAP analysis results for the PDML-GRU regression model at these two grid points are shown in Fig. 8. The SHAP values for the mean output directly reflect the magnitude of the features (introduced in Section 2.2) influencing the post-processed precipitation, while the SHAP values for the variance output indicate the uncertainty of the post-processed precipitation, i.e., the ensemble spread. At both grid points, the original precipitation forecasts make a positive contribution to both the mean and standard deviation of the post-processed precipitation outputs. When the original precipitation forecast is high, the PDML output tends to give larger mean and standard deviation values. It was also observed that the humidity forecast inputs have a significant impact on the PDML model's output. Similar to the original precipitation forecast, SH850 makes a positive contribution to both the mean and standard deviation of the output precipitation, while the contributions of SH850 and SH500 exhibit greater uncertainty, with performance differences observed at each level. For $level_1$ precipitation, temperature forecasts have a

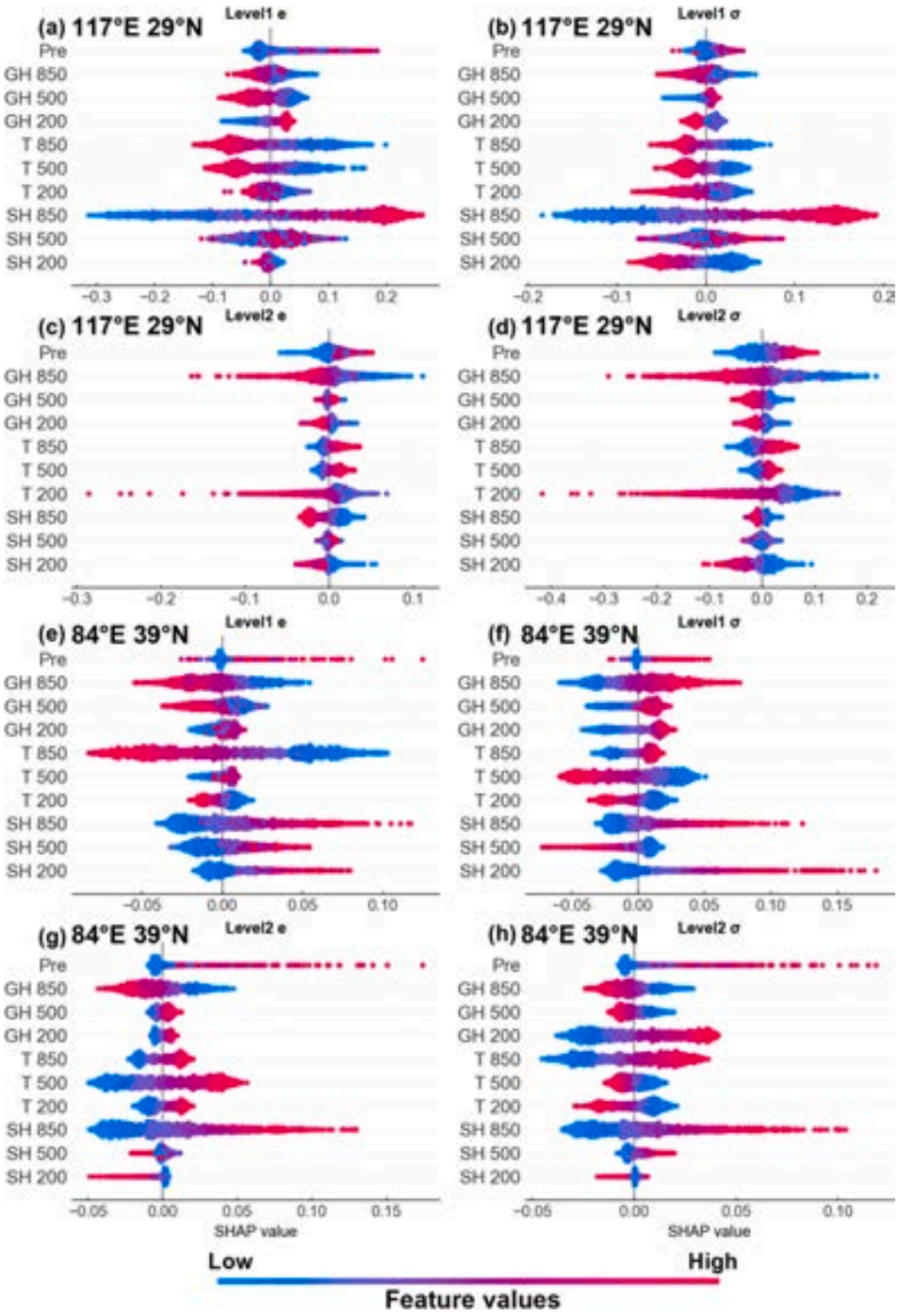



**Fig. 8.** SHAP values of each variable for the output of the PDML-GRU regression model at the grid points 117°E 29°N and 84°E 39°N. (a)-(d) display the SHAP values for the mean (e) and standard deviation (σ) of $\boldsymbol{level}_1$ precipitation output and $\boldsymbol{level}_2$ precipitation output at 117°E 29°N, respectively. (e)-(h) correspond to the SHAP values for the grid point at 84°E 39°N. The color mapping represents the magnitude of the feature values.

negative contribution to both the mean and standard deviation of the precipitation output, with a more prominent negative contribution observed for T850. For $level_2$ precipitation, the impact of temperature forecasts is relatively smaller. In the 117°E 29°N model, T200 has a clear negative contribution, whereas in the 84°E 39°N model, it shows a completely positive contribution to the mean output. The overall contribution of geopotential height is relatively small. For $level_1$ precipitation mean, GH850 and GH500 show a negative contribution, while GH200 has a positive contribution at both locations. For $level_2$ precipitation mean, GH850 is the only one with a noticeable negative contribution, and GH500 and GH200 have smaller contributions, with varying directions of impact.

Fig. S8 in the supplementary materials shows the SHAP analysis results for the classification model. We found that both the original precipitation forecast and humidity are the most significant features influencing the model. Both have a clear positive impact on $p_2$ and a negative impact on $p_0$. Additionally, for geopotential height and temperature, although the direction of influence varies across regions, a common characteristic is that the forecasts at 850 hPa and 500 hPa have a much stronger impact than those at 200 hPa. This is because 200 hPa is near the top of the troposphere, and its influence on precipitation formation is relatively minor.

### 5.1. Comparison with existing studies, limitations, and future work

Compared to existing DML models, our PDML approach extends its application scope from precipitation product correction to precipitation forecast post-processing, thereby demonstrating its feasibility in S2S forecast post-processing. Methodologically, we advance the framework from a binary classification task to a multiclass classification problem, while simultaneously transitioning from deterministic forecast outputs to ensemble forecast generation (Lei et al., 2022; Ling Zhang et al., 2021; Lyu and Yong, 2024; Xiao et al., 2022). This modification not only facilitates uncertainty quantification but also incorporates an additional regression task specifically designed for extreme precipitation events. To demonstrate the improvement of PDML compared to DML, we conducted an additional comparison using the LSTM model across 8 corresponding forecast periods. Considering that DML models can only generate deterministic forecasts, we selected only the deterministic evaluation metrics for the comparison. The results are shown in Fig. 9. After comparing metrics, we found that compared to DML, PDML showed significant improvements in all metrics at the daily scale, especially in CSI, which demonstrates that the additional incorporation of extreme precipitation classification enhances the ability to identify extreme precipitation events. At the accumulated scale, however, PDML did not yield any significant improvement over DML. We believe this may be due to the smoothing effect of the weekly sliding average, which simplified the mapping relationships, making more complex models prone to overfitting. However, despite the lack of additional improvements in deterministic metrics at this scale, PDML still holds value in quantifying uncertainty ranges, which is of significant importance for decision-makers in water resource management.

Our regression approach is analogous to that of Li et al.(2022) and Ji et al.(2022). Besides, we have omitted the censored parameters by introducing an additional classification task for no precipitation. Moreover, we place enhanced emphasis on extreme precipitation events. By explicitly addressing extreme events, the proposed approach enhances their detectability, which is further substantiated by the observed improvement in CSI. The multiclass approach also provides a clear direction for the further development of PDML. Due to limitations in both the scope of this paper and computational efficiency, only a three-class framework was employed. Future research will explore how to set up classifications and whether increasing the number of classes can further enhance performance.

From a model comparison perspective, we compared a wide range of deep learning (DL) models, including RNN-based, U-NET, MLP, and the most recent KAN model, in contrast to prior DML studies. Our findings reveal that both RNN-based and U-NET models perform the best within the PDML architecture. The RNN-based model is more suitable for daily precipitation forecasts, while the U-NET model excels in accumulated precipitation forecasts. We believe that this typical difference is caused by the timescale and the internal structure of the model. LSTM and GRU are trained on a grid-by-grid basis strategy, considering only the feature information of each grid and aggregating and correcting it from a timescale perspective. For daily scale data, forecast errors in time are more common, and therefore RNN-based models are better at post-processing such temporal biases. For accumulated scale data, since the data has inherently undergone at least a 7-day moving average, the differences in the data at each time step are small, and the information gain provided is limited. Additionally, time errors are diluted, preventing the RNN-based model from fully utilizing its time bias correction ability. Although the U-NET model does not consider temporal sequence information, it captures additional spatial information. Especially for accumulated scale data, where the time series provides limited information, spatial biases in the forecast (such as relative shifts in rainfall bands) are more likely. The U-NET model can effectively correct these spatial biases, something that the RNN-based model cannot address. This is the main reason why the U-NET model outperforms the RNN-based model on the accumulated scale. Besides, from the Sharpness metric, we observe that the ensemble divergence of the U-NET model is generally smaller, which is consistent across both daily scale and accumulated scale. This is actually related to the training method of the U-NET model. U-NET is a spatial convolutional model, and its optimization process is global, with the loss averaged across all grid points. In this context, although the eastern moist regions generally receive more

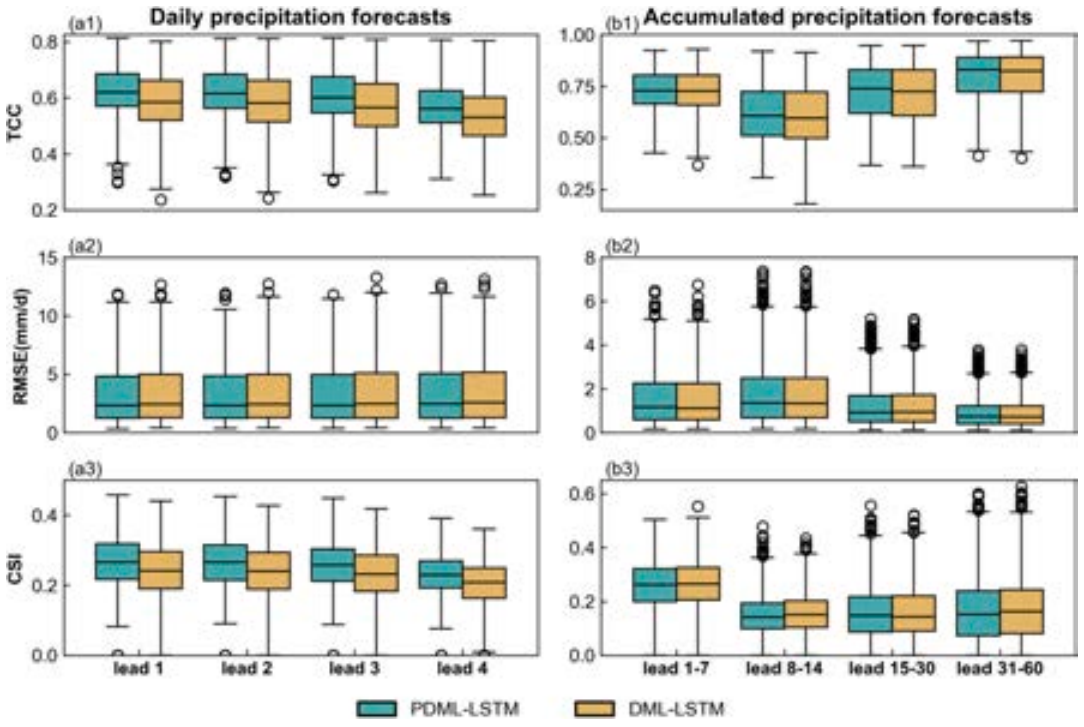



**Fig. 9.** Box plots of comparison of deterministic evaluation metrics between PDML-LSTM and DML-LSTM for each forecast period, with each value representing a grid point. (a1)-(b1) TCC, (a2)-(b2) RMSE, (a3)-(b3) CSI.

rainfall than the northwestern arid regions, their ensemble divergence is reduced during optimization to promote the overall loss decrease. This ensures global optimality rather than local optimality, resulting in smaller ensemble divergence for the U-NET model. In contrast, the RNN-based model follows a grid-point-wise training strategy, where each model only considers local optimization. This allows the eastern moist regions to generate larger ensemble divergences, while the northwestern arid regions produce smaller ensemble divergences, forming a clear distinction from the U-NET model. Fig. 10 illustrates the spatial distribution of $Dis_{95}$ on lead day 1. In the Bohai Sea region, the ensemble divergence formed by the RNN-based model is significantly greater than that of the U-NET model. On the accumulated scale, the RNN-based model also exhibits a decrease in ensemble divergence, which may be related to the limited information provided by the time series, as mentioned earlier. Both models outperform the state-of-the-art statistical model EPP in their suitable lead times. Additionally, while the KAN model has shown impressive performance in runoff forecasting, we observed that in the PDML framework, it does not surpass the more complex RNN-based and U-NET models when applied to precipitation forecasting (Granata et al., 2024). We suggest that replacing the fully connected layers with KAN layers in RNN-based and U-NET architectures could be a promising avenue for enhancing the original KAN model, leading to the development of RNN-KAN and U-NET-KAN models. Another potential direction is to combine RNN-based models with U-NET models, thereby capturing the relationships between features and targets from both spatial and temporal perspectives. Furthermore, the Conv-LSTM model has been successfully applied to radar precipitation forecasting tasks, and we believe future research could explore integrating it into the PDML framework (Shi et al., 2015).

One limitation in terms of model selection is that it does not consider the integration of different models within the PDML framework. The usage and validation of each PDML model are conducted independently. The classifiers and regressors in PDML can be a combination of different models, such as using RNN-based models that are better suited for classification tasks as classifiers, and U-Net models, which can capture spatial relationships, or other more advanced model architectures as regressors. Although this approach has been implemented in previous DML studies, there is no rigorous comparative experiment to demonstrate the effectiveness of such integration. Given that the main goal of this study is to propose the PDML framework, with the five commonly used DL models serving merely as baseline experiments to test the validity of PDML, we do not aim to achieve the most effective PDML model combination in this study. Testing the integration of models and other DL models will be explored in future research.

In terms of feature selection and model training, one of our limitations is the failure to account for the impact of climate change trends. Given the large temporal span of our study, covering 20 years, the influence of climate change on the mapping relationships cannot be ignored. In future research, we consider two potential approaches to address this issue. First, we could incorporate year and day of the year as additional auxiliary features into the PDML model, allowing the deep learning model to capture climate trends autonomously. Second, we could manually detrend the data before feeding it into the PDML model to remove the effects of climate change. However, we are also concerned that this approach might degrade data quality and introduce erroneous mappings.

The limited data availability is also a potential drawback of this method. As a sample-sensitive approach, deep learning methods, after classifying and dividing different sub-samples, lead to a reduction in the representativeness of the samples in each category, making it difficult for PDML to effectively capture the mapping relationship between forecasts and observations. This could pose challenges for regions with limited historical forecasting data. Additionally, the unreasonable threshold division can affect the quality of the data. Although we used a relative threshold method, the boundaries of the relative threshold may not be suitable for all regions. For example, in the South China region, due to the excessively humid climate conditions, the threshold we set may not be high enough, which could lead to unclear boundaries between extreme precipitation and general precipitation events, affecting proper identification. In the future, we will further explore methods to improve the threshold setting.

For feature importance analysis and model interpretability, we conducted SHAP analysis at two grid points, considering the computational cost. This approach has, to some extent, helped us understand the feature selection process and the internal logic of the model. However, the selected features are fixed, and there is a lack of comparative studies that explore other features and different combinations of features (Lin et al., 2023). Due to limitations in computational cost, we were unable to fully address this in the present study. But in future research, we believe that a more comprehensive interpretability analysis is essential, including more sophisticated feature selection and the exploration of their applicable spatial patterns.

Additionally, we compared the computation time of different PDML models relative to EPP, and the results are presented in Text. S3 and Fig. S9 of the supplementary materials. The results indicate that the PDML method has a significantly lower computation time compared to the EPP method, with the time required being more than an order of magnitude smaller. Among the PDML algorithms, the U-NET structure is the most time-efficient using only about 1/200 of computation time cost by EPP, as it is based on a convolutional architecture, where all sites are trained simultaneously and share a single model.

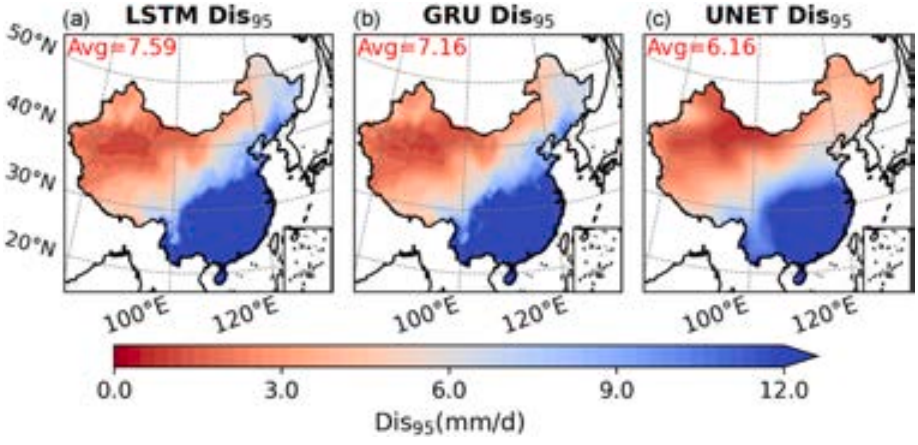



**Fig. 10.** Spatial distribution of the $Dis_{95}$ for lead day 1 precipitation forecasts. (a) and (b), (c) are derived from the LSTM, GRU, and U-NET models, respectively.

## 6. Conclusions

In this study, we propose a novel Probabilistic Double Machine Learning (PDML) method for S2S precipitation forecast post-processing and tested it under the new framework across eight different lead times, including four daily precipitation forecasts (lead day 1, 2, 3 and 4) and four accumulated precipitation forecasts (lead day 1–7, 8–14, 15–30 and 31–60). This new PDML architecture can quantify uncertainty by generating ensemble forecasts and enhance extreme precipitation forecasting ability by considering extreme precipitation events in both the classifier and regressor. We compared five different deep learning algorithms with one statistical algorithm, EPP. Interpretability and feature contribution analysis are also provided to gain a deeper understanding of the impact of each feature on the PDML. The key conclusions are as follows:

1) The PDML algorithm significantly improves the original forecasts for both deterministic forecasts and ensemble forecasts. On average, it achieves up to 85.8 % improvement in the TCC across all lead times and reduces RMSE by 45.3 %. In terms of extreme precipitation forecasting capabilities, the CSI (Critical Success Index) improves by an average of 294.6 %. For ensemble forecasts, the combined evaluation metrics of reliability and sharpness, measured by CRPS, show an 8.6 % reduction. When selecting the sophisticated model, the PDML model can also improve performance compared to the statistical EPP algorithm. At the daily scale, the improvement of PDML compared to DML is more significant.
2) From the comparison of DL models under PDML framework, for daily precipitation forecasts, we recommend using the RNN-based PDML model. For accumulated precipitation forecasts, the U-NET-based PDML model is preferred.
3) From the perspective of feature contribution and model interpretability, we found that, regardless of whether it is a classification or regression model, the original precipitation forecast and humidity forecast are the most influential input variables. Additionally, the importance of forecast variables at the 850 hPa and 500 hPa pressure levels is generally higher than that of forecast variables at the 200 hPa pressure level.
4) From the perspective of computational cost, the PDML model generally has a much lower computational cost than the EPP algorithm. The PDML-U-NET model, which has the lowest computational cost, requires only 1/200 of the computation time compared to the EPP algorithm.

## CRediT authorship contribution statement

**Shengsheng Zhan:** Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Aizhong Ye:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Lingyun Wu:** Writing – review & editing, Validation, Data curation. **Chenguang Zhao:** Writing – review & editing, Validation, Data curation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This study was supported by the National Key Research and Development Program of China (No. 2024YFF1306305), the Natural Science Foundation of China (No. 42171022), the BNU-FGS Global Environmental Change Program (No.2023-GC-ZYTS-06).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2025.133484.

## Data availability

Data will be made available on request.

## References

Chen, C., Zhang, Q., Kashani, M.H., Jun, C., Bateni, S.M., Band, S.S., Dash, S.S., Chau, K.-W., 2022. Forecast of rainfall distribution based on fixed sliding window long short-term memory. Eng. Appl. Comput. Fluid Mech. 16, 248–261. https://doi.org/10.1080/19942060.2021.2009374.

Chen, S., Feng, Y., Mao, Q., Li, H., Zhao, Y., Liu, J., Wang, H., Ma, D., 2024. Improving the accuracy of flood forecasting for Northeast China by the correction of global forecast rainfall based on deep learning. J. Hydrol. 640, 131733. https://doi.org/10.1016/j.jhydrol.2024.131733.

Duan, Q., Pappenberger, F., Wood, A., Cloke, H.L., Schaake, J.C. (Eds.), 2019. Handbook of Hydrometeorological Ensemble Forecasting. Springer Berlin Heidelberg, Berlin, Heidelberg. Doi: 10.1007/978-3-642-39925-1.

Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., Koch, B., 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. Remote Sens. Environ. 154, 102–114. https://doi.org/10.1016/j.rse.2014.07.028.

Gneiting, T., 2008. Editorial: probabilistic forecasting. J. R Stat. Soc. Ser. A Stat. Soc. 171, 319–321. https://doi.org/10.1111/j.1467-985X.2007.00522.x.

Granata, F., Zhu, S., Di Nunno, F., 2024. Advanced streamflow forecasting for Central European Rivers: The Cutting-Edge Kolmogorov-Arnold networks compared to Transformers. J. Hydrol. 645, 132175. https://doi.org/10.1016/j.jhydrol.2024.132175.

Herman, G.R., Schumacher, R.S., 2018. Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. Mon. Weather Rev. 146, 1571–1600. https://doi.org/10.1175/MWR-D-17-0250.1.

Huang, Z., Zhao, T., Xu, W., Cai, H., Wang, J., Zhang, Y., Liu, Z., Tian, Y., Yan, D., Chen, X., 2022. A seven-parameter Bernoulli-Gamma-Gaussian model to calibrate subseasonal to seasonal precipitation forecasts. J. Hydrol. 610, 127896. https://doi.org/10.1016/j.jhydrol.2022.127896.

Jahangir, M.S., Quilty, J., 2024. Generative deep learning for probabilistic streamflow forecasting: Conditional variational auto-encoder. J. Hydrol. 629, 130498. https://doi.org/10.1016/j.jhydrol.2023.130498.

Ji, Y., Zhi, X., Ji, L., Zhang, Y., Hao, C., Peng, T., 2022. Deep-learning-based post-processing for probabilistic precipitation forecasting. Front. Earth Sci. 10, 978041. https://doi.org/10.3389/feart.2022.978041.

Jolliffe, I.T., Stephenson, D.B. (Eds.), 2011. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley, Chichester. https://doi.org/10.1002/9781119960003.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrol. Earth Syst. Sci. 26, 1673–1693. https://doi.org/10.5194/hess-26-1673-2022.

Kolachian, R., Saghafian, B., 2019. Deterministic and probabilistic evaluation of raw and post processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes. Theor. Appl. Climatol. 137, 1479–1493. https://doi.org/10.1007/s00704-018-2680-5.

Kossieris, P., Tsoukalas, I., Brocca, L., Mosaffa, H., Makropoulos, C., Anghelea, A., 2024. Precipitation data merging via machine learning: Revisiting conceptual and technical aspects. J. Hydrol. 637, 131424. https://doi.org/10.1016/j.jhydrol.2024.131424.

Laio, F., Tamea, S., 2007. Verificationtools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci. 11, 1267–1277. https://doi.org/10.5194/hess-11-1267-2007.

Larraondo, P.R., Renzullo, L.J., Van Dijk, A.I.J.M., Inza, I., Lozano, J.A., 2020. Optimization of deep learning precipitation models using categorical binary metrics. J. Adv. Model. Earth Syst. 12, e2019MS001909. https://doi.org/10.1029/2019MS001909.

Lei, H., Zhao, H., Ao, T., 2022. A two-step merging strategy for incorporating multi-source precipitation products and gauge observations using machine learning classification and regression over China. Hydrol. Earth Syst. Sci. 26, 2969–2995. https://doi.org/10.5194/hess-26-2969-2022.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., Di, Z., 2017. A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. Wires Water 4, e1246.

Li, W., Duan, Q., Ye, A., Miao, C., 2019. An improved meta-Gaussian distribution model for post-processing of precipitation forecasts by censored maximum likelihood estimation. J. Hydrol. 574, 801–810. https://doi.org/10.1016/j.jhydrol.2019.04.073.

Li, W., Kiaghadi, A., Dawson, C., 2021. Exploring the best sequence LSTM modeling architecture for flood prediction. Neural Comput. Appl. 33, 5571–5580. https://doi.org/10.1007/s00521-020-05334-3.

Li, W., Pan, B., Xia, J., Duan, Q., 2022. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. J. Hydrol. 605, 127301. https://doi.org/10.1016/j.jhydrol.2021.127301.

Li, X., Wu, H., Nanding, N., Chen, S., Hu, Y., Li, L., 2023. Statistical bias correction of precipitation forecasts based on quantile mapping on the sub-seasonal to seasonal scale. Remote Sens. 15, 1743. https://doi.org/10.3390/rs15071743.

Lin, X., Fan, J., Hou, Z.J., Wang, J., 2023. Machine learning of key variables impacting extreme precipitation in various regions of the contiguous United States. J. Adv. Model. Earth Syst. 15, e2022MS003334. https://doi.org/10.1029/2022MS003334.

Liu, S., Li, W., Duan, Q., 2023. Spatiotemporal variations in precipitation forecasting skill of three global subseasonal prediction products over China. J. Hydrometeorol. 24, 2075–2090. https://doi.org/10.1175/JHM-D-23-0071.1.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M., 2024. KAN: Kolmogorov-Arnold Networks. arXiv: 2404.19756. Doi: 10.48550/arXiv.2404.19756.

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. arXiv preprint. arXiv:1705.07874. Doi: 10.48550/arXiv.1705.07874.

Luo, Z., Zhang, S., Shao, Q., Wang, L., Wang, S., Wang, L., 2024. A new method to improve precipitation estimates by blending multiple satellite/reanalysis-based precipitation products and considering observations and terrestrial water budget balance. J. Hydrol. 635, 131188. https://doi.org/10.1016/j.jhydrol.2024.131188.

Lyu, Y., Yong, B., 2024. A Novel Double Machine Learning Strategy for Producing High-Precision Multi-Source Merging Precipitation Estimates Over the Tibetan Plateau. Water Resour. Res. 60, e2023WR035643. https://doi.org/10.1029/2023WR035643.

Lyu, Y., Zhu, S., Zhi, X., Ji, Y., Fan, Y., Dong, F., 2023. Improving subseasonal-to-seasonal prediction of summer extreme precipitation over southern china based on a deep learning method. Geophys. Res. Lett. 50, e2023GL106245. https://doi.org/10.1029/2023GL106245.

Lyu, Y., Zhu, S., Zhi, X., Wang, J., Ji, Y., Fan, Y., Dong, F., 2024. Significant advancement in subseasonal-to-seasonal summer precipitation ensemble forecast skills in China mainland through an innovative hybrid CSG-UNET method. Environ. Res. Lett. 19, 074055. https://doi.org/10.1088/1748-9326/ad5577.

Manzanas, R., Gutiérrez, J.M., Bhend, J., Hemri, S., Doblas-Reyes, F.J., Torralba, V., Penabad, E., Brookshaw, A., 2019. Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset. Clim. Dyn. 53, 1287–1305. https://doi.org/10.1007/s00382-019-04640-4.

Mao, Y., Sorteberg, A., 2020. Improving radar-based precipitation nowcasts with machine learning using an approach based on Random Forest. Weather Forecast. 35, 2461–2478. https://doi.org/10.1175/WAF-D-20-0080.1.

Martinez-Villalobos, C., Neelin, J.D., 2019. Why Do precipitation intensities tend to follow gamma distributions? J. Atmospheric Sci. 76, 3611–3631. https://doi.org/10.1175/JAS-D-18-0343.1.

Moghaddam, D.D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A.T., Nalivan, O.A., Tien Bui, D., 2020. The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. CATENA 187, 104421. https://doi.org/10.1016/j.catena.2019.104421.

Ni, L., Wang, D., Singh, V.P., Wu, J., Wang, Y., Tao, Y., Zhang, J., 2020. Streamflow and rainfall forecasting by two long short-term memory-based models. J. Hydrol. 583, 124296. https://doi.org/10.1016/j.jhydrol.2019.124296.

Oliveira, E.C.L.D., Nogueira Neto, A.V., Santos, A.P.P.D., Da Costa, C.P.W., Freitas, J.C.G. D., Souza-Filho, P.W.M., Rocha, R.D.L., Alves, R.C., Franco, V.D.S., Carvalho, E.C.D., Tedeschi, R.G., 2023. Precipitation forecasting: from geophysical aspects to machine learning applications. Front. Clim. 5, 1250201. https://doi.org/10.3389/fclim.2023.1250201.

Ortiz-García, E.G., Salcedo-Sanz, S., Casanova-Mateo, C., 2014. Accurate precipitation prediction with support vector classifiers: a study including novel predictive variables and observational data. Atmospheric Res. 139, 128–136. https://doi.org/10.1016/j.atmosres.2014.01.012.

Rivoire, P., Martius, O., Naveau, P., Tuel, A., 2023. Assessment of subseasonal-to-seasonal (S2S) ensemble extreme precipitation forecast skill over Europe. Nat. Hazards Earth Syst. Sci. 23, 2857–2871. https://doi.org/10.5194/nhess-23-2857-2023.

Scheuerer, M., Hamill, T.M., 2015. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. Mon. Weather Rev. 143, 4578–4596. https://doi.org/10.1175/MWR-D-15-0061.1.

Scheuerer, M., Switanek, M.B., Worsnop, R.P., Hamill, T.M., 2020. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. Mon. Weather Rev. 148, 3489–3506. https://doi.org/10.1175/MWR-D-20-0096.1.

Senocak, A.U.G., Yilmaz, M.T., Kalkan, S., Yucel, I., Amjad, M., 2023. An explainable two-stage machine learning approach for precipitation forecast. J. Hydrol. 627, 130375. https://doi.org/10.1016/j.jhydrol.2023.130375.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., Woo, W., 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv preprint. arXiv:1506.04214. Doi: 10.48550/arXiv.1506.04214.

Specq, D., Batté, L., 2020. Improving subseasonal precipitation forecasts through a statistical–dynamical approach : application to the southwest tropical Pacific. Clim. Dyn. 55, 1913–1927. https://doi.org/10.1007/s00382-020-05355-7.

Tao, Y., Duan, Q., Ye, A., Gong, W., Di, Z., Xiao, M., Hsu, K., 2014. An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin. J. Hydrol. 519, 2890–2905. https://doi.org/10.1016/j.jhydrol.2014.04.040.

Tripathy, K.P., Mishra, A.K., 2024. Deep learning in hydrology and water resources disciplines: concepts, methods, applications, and research directions. J. Hydrol. 628, 130458. https://doi.org/10.1016/j.jhydrol.2023.130458.

Vitart, F., 2014. Evolution of ECMWF sub-seasonal forecast skill scores: Evolution of the ECMWF Sub-Seasonal Forecast Skill. Q. J. r. Meteorol. Soc. 140, 1889–1899. https://doi.org/10.1002/qj.2256.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A.W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., Zhang, L., 2017. The Subseasonal to Seasonal (S2S) prediction project database. Bull. Am. Meteorol. Soc. 98, 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1.

Vitart, F., Robertson, A.W., Spring, A., Pinault, F., Roškar, R., Cao, W., Bech, S., Bienkowski, A., Caltabiano, N., De Coning, E., Denis, B., Dirkson, A., Dramsch, J., Dueben, P., Gierschendorf, J., Kim, H.S., Nowak, K., Landry, D., Lledó, L., Palma, L., Rasp, S., Zhou, S., 2022. Outcomes of the WMO prize challenge to improve subseasonal to seasonal predictions using artificial intelligence. Bull. Am. Meteorol. Soc. 103, E2878–E2886. https://doi.org/10.1175/BAMS-D-22-0046.1.

Wang, B., Huang, F., Wu, Z., Yang, J., Fu, X., Kikuchi, K., 2009. Multi-scale climate variability of the South China Sea monsoon: a review. Dyn. Atmospheres Oceans 47, 15–37. https://doi.org/10.1016/j.dynatmoce.2008.09.004.

Wang, Q.J., 2001. A Bayesian Joint Probability Approach for flood record augmentation. Water Resour. Res. 37, 1707–1712. https://doi.org/10.1029/2000WR900401.

Weyn, J.A., Durran, D.R., Caruana, R., 2020. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. J. Adv. Model. Earth Syst. 12, e2020MS002109. https://doi.org/10.1029/2020MS002109.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J.T., Lazo, J.K., Kumar, A., Vitart, F., Coughlan De Perez, E., Ray, A.J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A.P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N.J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T.J., Street, R., Jones, L., Remenyi, T.A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., Zebiak, S.E., 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol. Appl. 24, 315–325. https://doi.org/10.1002/met.1654.

White, C.J., Domeisen, D.I.V., Acharya, N., Adefisan, E.A., Anderson, M.L., Aura, S., Balogun, A.A., Bertram, D., Bluhm, S., Brayshaw, D.J., Browell, J., Büeler, D., Charlton-Perez, A., Chourio, X., Christel, I., Coelho, C.A.S., DeFlorio, M.J., Delle Monache, L., Di Giuseppe, F., García-Solórzano, A.M., Gibson, P.B., Goddard, L., González Romero, C., Graham, R.J., Graham, R.M., Grams, C.M., Halford, A., Huang, W.T.K., Jensen, K., Kilavi, M., Lawal, K.A., Lee, R.W., MacLeod, D., Manrique-Suñén, A., Martins, E.S.P.R., Maxwell, C.J., Merryfield, W.J., Muñoz, Á.G., Olaniyan, E., Otieno, G., Oyedepo, J.A., Palma, L., Pechlivanidis, I.G., Pons, D., Ralph, F.M., Reis, D.S., Remenyi, T.A., Risbey, J.S., Robertson, D.J.C., Robertson, A. W., Smith, S., Soret, A., Sun, T., Todd, M.C., Tozer, C.R., Vasconcelos, F.C., Vigo, I., Waliser, D.E., Wetterhall, F., Wilson, R.G., 2022. Advances in the application and utility of subseasonal-to-seasonal predictions. Bull. Am. Meteorol. Soc. 103, E1448–E1472. https://doi.org/10.1175/BAMS-D-20-0224.1.

Wu, J., Gao, X., Giorgi, F., Chen, D., 2017. Changes of effective temperature and cold/hot days in late decades over China based on a high resolution gridded observation dataset. Int. J. Climatol. 37, 788–800. https://doi.org/10.1002/joc.5038.

Xiao, S., Zou, L., Xia, J., Yang, Z., Yao, T., 2022. Bias correction framework for satellite precipitation products using a rain/no rain discriminative model. Sci. Total Environ. 818, 151679. https://doi.org/10.1016/j.scitotenv.2021.151679.

Xu, Y., Gao, X., Shen, Y., Xu, C., Shi, Y., Giorgi, F., 2009. A daily temperature dataset over China and its application in validating a RCM simulation. Adv. Atmospheric Sci. 26, 763–772. https://doi.org/10.1007/s00376-009-9029-z.

Xu, Y., Tang, G., Li, L., Wan, W., 2024. Multi-source precipitation estimation using machine learning: clarification and benchmarking. J. Hydrol. 635, 131195. https://doi.org/10.1016/j.jhydrol.2024.131195.

Ye, A., Deng, X., Ma, F., Duan, Q., Zhou, Z., Du, C., 2017. Integrating weather and climate predictions for seamless hydrologic ensemble forecasting: a case study in the Yalong River basin. J. Hydrol. 547, 196–207. https://doi.org/10.1016/j.jhydrol.2017.01.053.

Yin, G., Yoshikane, T., Kaneko, R., Yoshimura, K., 2023. Improving global subseasonal to seasonal precipitation forecasts using a support vector machine-based method. J. Geophys. Res. Atmospheres 128, e2023JD038929. https://doi.org/10.1029/2023JD038929.

Zhang, Y., Ye, A., 2021. Machine learning for precipitation forecasts post-processing — Multi-model comparison and experimental investigation. J. Hydrometeorol. https://doi.org/10.1175/JHM-D-21-0096.1.

Zhang, Y., Ye, A., Nguyen, P., Analui, B., Sorooshian, S., Hsu, K., 2022. QRF4P-NRT: probabilistic post-processing of near-real-time satellite precipitation estimates using Quantile Regression Forests. Water Resour. Res. 58, e2022WR032117. https://doi.org/10.1029/2022WR032117.