Geosci. Model Dev., 19, 6027–6041, 2026
https://doi.org/10.5194/gmd-19-6027-2026

Geoscientific Model Development
Open Access
EGU

A self-supervised precipitation forecast verification based on contrastive learning

Yanwen Wang[1,2,★], **Shuwen Huang**[1,★], **Qian Li**[1,2], **Xuan Peng**[1,2], **Haoming Chen**[3], **Kefeng Zhu**[4], **Liwen Wang**[1], **and Sheng Li**[1]

[1]College of Meteorology and Oceanography, National University of Defense Technology, Changsha, 410072, China
[2]High Impact Weather Key Laboratory of CMA, Changsha, 410072, China
[3]State Key Laboratory of Severe Weather Meteorological Science and Technology, Chinese Academy of Meteorological Sciences, Beijing, 100081, China
[4]Nanjing Innovation Institute for Atmospheric Sciences, Chinese Academy of Meteorological Sciences-Jiangsu Meteorological Service, Nanjing, 210041, China
★These authors contributed equally to this work.

Correspondence: Yanwen Wang (feixian_wangyw@qq.com)

Received: 19 November 2025 – Discussion started: 23 February 2026
Revised: 15 May 2026 – Accepted: 17 June 2026 – Published: 8 July 2026

Abstract. Accurate precipitation forecast verification (PFV) is essential for improving forecasting models and supporting disaster management. However, current PFV methods remain limited, point-to-point matching is overly sensitive to minor errors, whereas spatial verification typically necessitates parameter tuning or heuristic rules derived from expert knowledge, which constrains their availability. To tackle these issues, we are inspired by the success of deep learning in image verification through extracting high-level features, and thus propose a self-supervised contrastive learning-based PFV method (CLPFV). First, CLPFV uses precipitations augmentation (displacement, intensity, area) to simulate actual forecast errors and construct positive and negative training sample pairs. Subsequently, with a novel loss function proportionally penalizing forecast errors, a backbone network is trained in CLPFV to extract high-level precipitation features. Finally, the cosine similarity of features is calculated as CLPFV's verification score. Experiments demonstrate that CLPFV outperforms traditional and spatial verifications in different degrees of forecast errors and aligns better with expert assessments. In general, CLPFV offers an efficient deep learning solution for PFV tasks.

1 Introduction

Precipitation forecasting is crucial for many practical applications such as transportation management, agricultural production, and disaster mitigation (Saavedra Valeriano et al., 2010). However, due to the complexity of precipitation formation and its numerous influencing factors, precipitation forecasting remains one of the most challenging meteorological issues (Chen et al., 2024), whilst precipitation forecast verification (PFV) is also a crucial step in the corresponding studies (Zhu et al., 2022). Therefore, developing a reliable PFV method is essential for understanding forecasting quality and promoting forecast model improvements (Xu et al., 2021), because reliable verifications could accurately reflect the specific degree of forecast error, thereby providing targeted guidance for refining precipitation forecasting techniques (Lee et al., 2011; Dorninger et al., 2020).

Precipitation forecast error, i.e., the deviation between forecasted and observed precipitations, mainly manifests in three aspects: (1) displacement error, i.e., the deviation in spatial positions, (2) intensity error, i.e., the deviation in precipitation intensity, (3) area size error, i.e., the deviation in area. Therefore, the PFV needs to comprehensively account for these three aspects to ensure the reliability of its verified forecast errors (Ebert and Gallus, 2009; Li et al., 2024). The existing PFV methods are usually catego-

rized into three classes: expert manual verification methods, traditional point-to-point verification methods, and spatial verification methods (Cassola et al., 2015). Expert manual verification methods primarily rely on experts' professional knowledge to carry out the verification through visual inspections. Although this kind of method is straightforward and expert-endorsed, it is labor-intensive and time-consuming, susceptible to subjective bias, restricting large-scale applications.

To improve verifying efficiency, point-to-point verification methods were developed and have been widely used (Rossa et al., 2008). They automatically compare forecasted and observed precipitations at matching grids, generating verification scores such as probability of detection (POD), false alarm ratio (FAR), and threat score (TS) to quantify forecast accuracy. Traditional point-to-point methods can reliably differentiate between the completely accurate precipitation forecasting and significantly deviated one via their verification scores. However, their grid-to-grid matching strictly makes them overly sensitive to minor forecast errors, especially in high-resolution precipitation forecasting tasks (van der Plas et al., 2017; Gofa et al., 2022). For example, traditional point-to-point verification methods frequently encounter the "double penalty" problem, where minor displacement errors receive double penalties (both False Alarms and Misses), consequently leading to underestimated forecast accuracy (Jain et al., 2023).

To mitigate the oversensitivity of verification to minor errors, researchers have developed a series of spatial verification methods (Gofa et al., 2018), among which neighborhood methods and object-based methods are representative (Gilleland et al., 2009, 2010; Dorninger et al., 2018). Neighborhood methods introduce neighbor buffering operations to the grids of forecasted precipitations in the verification process, thereby contributing to tolerance of minor displacement errors (Roberts and Lean, 2008; Mittermaier and Roberts, 2010). Object-based methods verify forecasted precipitations with grids-based statistical metrics as their object indicators instead of matching their individual grids to generate verification scores (Davis et al., 2006; Wernli et al., 2008, 2009). Although spatial verification methods can achieve better precipitation forecast verifications, their effectiveness often relies on appropriate parameters and rules tailored to the specific task, which typically require extensive expert knowledge or numerous pre-experiments. In practice, however, these requirements are often difficult to meet, significantly limiting the availability of spatial verification methods. More notably, if the parameters and rules are improperly set, spatial verification methods may not only be inaccurate enough but could even underperform traditional point-to-point methods. Moreover, current spatial verification methods still struggle to effectively handle inappropriate verifications of intensity and area size errors. Therefore, a more comprehensive verification method is needed to tolerate minor errors yet penalize significant ones, thereby better reflecting the varying degrees of error.

The limitation of existing spatial verification methods essentially stems from their reliance on predefined parameters and rules, preventing them from truly capturing the spatial distributions of observed and forecast precipitation fields. Consequently, conducting PFV from an overall structural perspective promises more reliable results. Inspired by this, we propose a deep-learning-based PFV method that evaluates forecast performance by comparing the overall high-level features of observed and forecasted precipitations. This approach leverages the exceptional capabilities of deep learning in simulating human cognitive processes and extracting complex features, as well as its remarkable success in image verification practices in recent years (Gidaris et al., 2018; Oord et al., 2018; Qiu et al., 2023). Furthermore, it is particularly important to note that deep learning models typically require labeled training data, but the costs of labeling and the complexity of precipitation forecasting tasks have limited the number of available labeled samples. In other words, addressing the issue of lacking labeled samples properly is the key to ensuring the practical applicability of verification based on deep learning. In sight of this, self-supervised learning (SSL) is a well-suited approach, which uses pretext tasks to generate supervised signals from unlabeled data for implementing the learning process (Liu et al., 2021). Among various SSL models, contrastive learning (CL) suits the precipitation forecast verification tasks well, which can automatically learn the representative features from the data by comparing the similarities and differences between augmented and different unlabeled samples (Saunshi et al., 2019; Xu et al., 2025), achieving competitive verifications without the need for labeled samples (He et al., 2016; Chen et al., 2020; Grill et al., 2020).

Based on the above considerations, this paper proposes a self-supervised precipitation forecast verification method based on contrastive learning (CLPFV), which can train deep neural networks to extract high-level abstract features for verification without the need of labeled data, and reflecting different degrees of forecast errors by multi augmentations and improved loss function. The remainder of this paper is organized as follows. Section 2 introduces the proposed method. Next, we discuss the experiments and results in Sect. 3. Finally, Sect. 4 provides a summary and conclusion.

2 Methodology

In this section, we present the proposed verification method, named CLPFV, in detail. The core idea of CLPFV is to conduct the verification by shifting from grid-matching to an overall high-level structural similarity comparison through self-supervised contrastive learning. To be more specific, we first used multi-dimensional precipitation augmentations in CLPFV to create intrinsic supervisory signals from unlabeled

data to address the scarcity of labeled samples. Subsequently, we designed an improved contrastive loss function that applies proportional penalties to forecast errors when extracting high-level precipitation features, thereby reasonably reflecting the gradient of errors. Finally, the result is directly calculated by comparing the high-dimensional features of observed and forecasted precipitations, achieving the verification from an overall structural perspective.

Following this basic idea, the framework of CLPFV consists of three stages as illustrated in Fig. 1: precipitations augmentation, features extraction, verification score calculation. In the first stage, the original precipitations are multi augmented in displacement, intensity, and area size to simulate realistic precipitation forecast errors with different degrees, addressing the problem of lacking labeled samples noted in the introduction, where the multi augmented precipitations and their corresponding original precipitations form positive sample pairs, while different original precipitations serve as negative pairs. In the stage of features extraction, based on these sample pairs, a deep learning network, ResNet-18 adopted in this research, is trained through the improved contrastive loss function to extract high-level abstract features of precipitation. The above two stages constitute the training process of CLPFV. As to the practical application of CLPFV (i.e., the final stage of verification score calculation), the forecasted and observed precipitations will be input into the trained deep learning network to obtain their features, and then the feature-based similarity is calculated as the verification score. The specific steps of CLPFV are introduced in the following contents.

2.1 Precipitations Augmentation

As Li et al. (2024) stated, precipitation forecast errors can be primarily manifested by three aspects: displacement, intensity, and area size. While these three aspects fall short of providing a comprehensive depiction of the spatial morphology and structure of precipitation, they offer a straightforward and quantitative assessment of forecast errors and are commonly used in PFV studies (Ebert and Gallus, 2009). Therefore, the augmentations of original precipitations in CLPFV should be accordingly set as Fig. 2 illustrated. Notably, to simulate the varying degrees of forecast errors in real-world precipitation forecasting, we apply multi augmentations with different magnitudes to the original precipitations, rather than only using a single augmentation.

Given that the precipitation field data are raster data, any precipitation field can be described as a set of its contained precipitation grids x_k. Then Eq. (1) presents the matrix expression of x_k, where i is the row number of x_k, j is the column number of x_k, and p is the precipitation intensity value of x_k. In this way, the precipitation augmentation process can be expressed as a transformation operation performed on x_k with a matrix $\mathbf{M}$. All transformed grids x'_k form the augmented precipitation. By controlling the values of its elements, transformation matrix $\mathbf{M}$ can simulate any specified degree of three kinds of precipitation errors, laying the foundation for conducting comprehensive experiments with diverse scenarios of precipitation forecasting.

$$x_k = \begin{bmatrix} i \\ j \\ p \end{bmatrix} \tag{1}$$

Displacement augmentation is to mimic displacement errors of forecasted precipitations. Transformation matrix $\mathbf{M}$ here is expressed by Eq. (2), where d_i and d_j represent the offsets of the precipitation grid in the latitude (row) and longitude (column) directions respectively, and displacement error can be calculated as $\sqrt{d_i^2 + d_j^2}$. Displacement augmentation can be expressed as adding x_k and $\mathbf{M}$ shown in Eq. (3).

$$\mathbf{M} = \begin{bmatrix} d_i \\ d_j \\ 0 \end{bmatrix} \tag{2}$$

$$x'_k = x_k + \mathbf{M} = \begin{bmatrix} i + d_i \\ j + d_j \\ p \end{bmatrix} \tag{3}$$

Intensity augmentation is designed to simulate precipitation intensity errors of forecasted precipitations, where increasing or decreasing every grid's precipitation intensity value to a certain extent. Transformation matrix $\mathbf{M}$ here is expressed by Eq. (4), where t denotes the proportion of precipitation increase/decrease, and then intensity augmentation can be expressed as Eq. (5).

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1+t \end{bmatrix} \tag{4}$$

$$x'_k = \mathbf{M} \times x_k = \begin{bmatrix} i \\ j \\ p(1+t) \end{bmatrix} \tag{5}$$

Precipitation area size augmentation is designed to simulate area size errors of forecasted precipitations, by scaling the precipitation region to a certain proportion. Scaling is performed around the centroid (i_c, j_c) of the precipitation grids with a scaling percentage denoted s, and it consists of three steps. First, move the precipitation region until its centroid arrives at the coordinate origin $(0, 0)$. Next, scale the moved precipitation region. Finally, move the scaled precipitation region back until its centroid returns to (i_c, j_c). Hence, there are three transformation matrixes (i.e., $\mathbf{M}_1$, $\mathbf{M}_2$, and $\mathbf{M}_3$ shown in Eq. 6) used here, and the scaling transformed grid x'_k can be expressed by Eq. (7). Since scaling will cause changes in the number and positions of precipitation grids, it is necessary to interpolate the scaled grids. Specifically, when the precipitation is enlarged, gaps will appear between

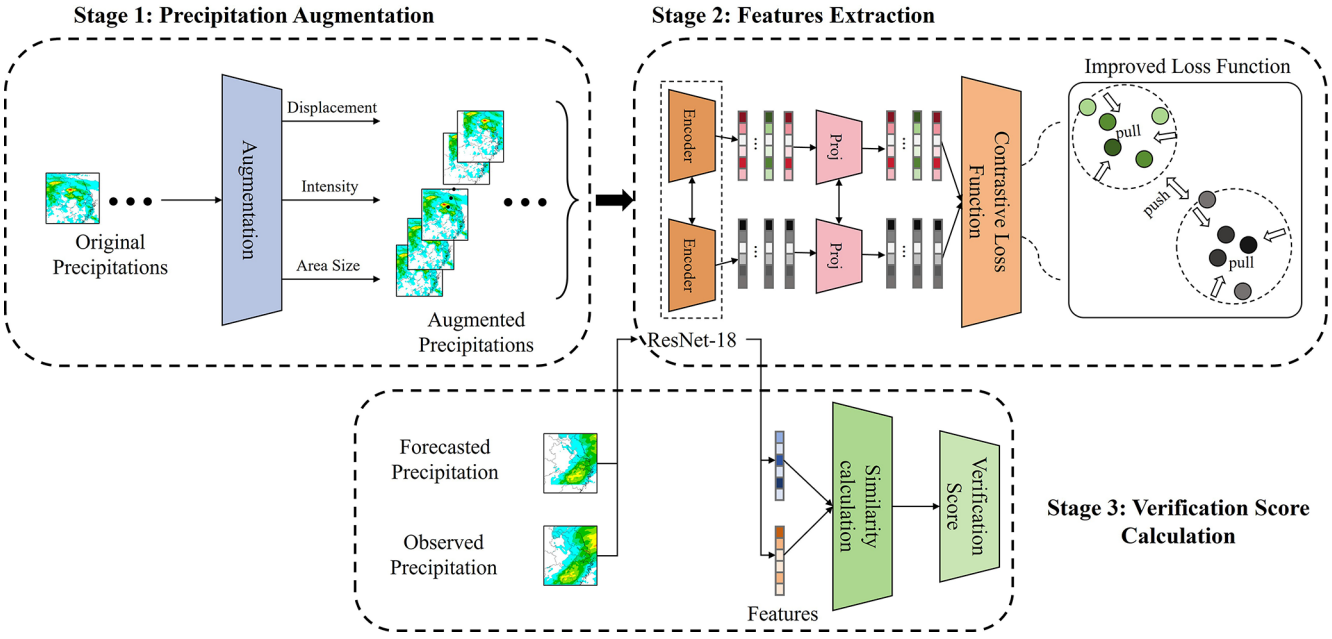


Figure 1. The overall framework of CLPFV.

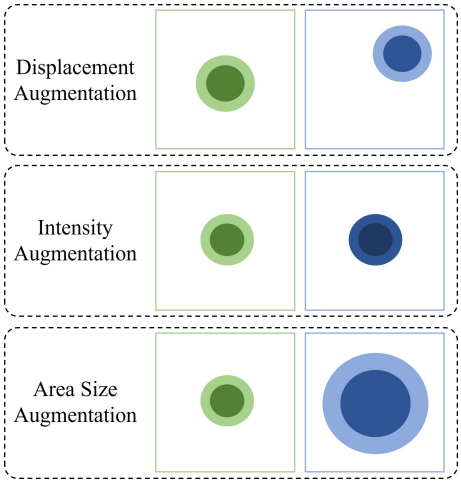


Figure 2. Visualizations of displacement augmentation, intensity augmentation, and area size augmentation. For each row, left column represents the original precipitations, right column represents the augmented precipitations, with outer circle coverage denoting precipitation area size and colour shading indicating precipitation intensity.

the scaled grids, and it will be filled with bilinear interpolation. When the precipitation is reduced, there will be multiple scaled grids at the same coordinate, and we average their precipitation intensities to obtain the final value of that coordinate.

$$\mathbf{M}_1 = \begin{bmatrix} -i_c \\ -j_c \\ 0 \end{bmatrix}, \quad \mathbf{M}_2 = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{M}_3 = \begin{bmatrix} i_c \\ j_c \\ 0 \end{bmatrix} \tag{6}$$

$$x_k^s = (\mathbf{M}_2 \times (x_k + \mathbf{M}_1)) + \mathbf{M}_3 = \begin{bmatrix} s\,(i - i_c) + i_c \\ s\,(j - j_c) + j_c \\ p \end{bmatrix} \tag{7}$$

These augmented precipitations, along with the original ones, form positive sample pairs, while different original precipitations forming negative sample pairs. Together, these positive and negative sample pairs provide the training data for the next stage of features extraction. By precipitations augmentation and features extraction, CLPFV can train models using original precipitations samples even without artificial labels, and finally conduct automated precipitation forecast verification.

2.2 Features extraction

In the stage of features extraction, an encoder is constructed to convert precipitations data (in raster format) into high-level abstract features that capture key patterns, which enables the next stage to calculate verification scores by quantifying the feature-based similarity between forecasted and observed precipitations. Two components are critical in this encoder: the backbone network, which converts raster data

into feature vectors, and the loss function, which optimizes the representativeness of high-level abstract features.

In this research, we employ ResNet-18 as the backbone network because it can effectively alleviate the vanishing gradient issue and has the advantage of being lightweight (He et al., 2016). As Fig. 3 illustrated, ResNet18 starts initial features extraction of precipitation raster data via a convolutional kernel of 7×7 size and a stride of 2. Subsequently, a max pooling layer also with a stride of 2 reduces the size of the feature maps, thereby decreasing the complexity of the subsequent computations. The feature maps are then input into four residual blocks, each containing two 3×3 convolutional layers and one 1×1 convolutional layer. In these four residual blocks, the number of channels in the 3×3 convolutional layers is set to 64, 128, 256, and 512 in succession. This incremental setup facilitates the network in progressively extracting richer information about the precipitation raster data. The presence of the 1×1 convolutional layer effectively alleviates the vanishing gradient problem commonly encountered in the networks. After the residual blocks, an average pooling layer with a stride of 2 is employed to further reduce the size of the feature maps. Finally, these features are connected to a fully connected layer, generating a high-level features vector that captures of the precipitation raster data. To minimize feature redundancy and improve their representation, a projection head is added to the end of backbone network based on a multilayer perceptron structure (Chen et al., 2020), which is responsible for reducing the features' dimensionality output by the backbone network. Specifically, the projection headfirst takes the backbone network's output features vector as input, applies a linear transformation through a fully connected layer, and then uses an activation function to introduce non-linearity into the features, ultimately outputting a low-dimensional features vector.

After determining to employ ResNet-18 as the backbone network, it is necessary to use appropriate loss functions in the training process to guide the network optimization. Since conventional contrastive learning verification tasks only need to determine whether data belongs to a specific category or identify its class membership, their loss functions can just pull positive samples closer while push negative samples apart in the features space, e.g., InfoNCE (Oord et al., 2018), one of the mostly used loss functions. However, CLPFV is designed to reflect different degrees of forecast errors by its verification scores, which demands the loss function not only to distinguish between positive and negative samples but also to differentiate positive samples with varying tiers (i.e., precipitations with multi augmentations) to reflect the error degrees. To achieve this, we represent an improved loss function by connecting a penalty term with InfoNCE. As illustrated in Fig. 4, this improved loss function maintains the capability of pulling positive samples closer and pushing negative samples apart, moreover, it additionally introduces corresponding penalties based on the augmentation scale of precipitations.

Equation (8) shows CLPFV's loss function in detail, where f_i is the ith original precipitation, f_+ is its augmented precipitation (i.e., its positive sample), and f_- is another original precipitation (i.e., its negative sample). All of them (f_i, f_+, and f_-) are temporary extracted features of corresponding precipitations in the training process. $p(f_+)$ is the punishment function of augmented precipitations. Equation (9) shows the specific equation of punishment function $p(f_+)$, which is the combination of the magnitude square of all kinds augmentations m with a penalty coefficient λ. Since CLPFV should tolerate minor forecast errors while reflecting significant ones, the greater the magnitude of augmentation, the stronger the penalty will be. Additionally, since both intensity and area size augmentation can be set in either positive or negative directions, a quadratic function is suitable and is therefore adopted as $p(f_+)$ in this research. Specifically, the quadratic function provides a smooth and differentiable penalty for model training and features extraction. It ensures the forecast verification remains tolerant of minor errors while applying increasingly strict penalties to larger errors, effectively distinguishing different ties of forecast quality.

$$L_{\mathrm{CLPV}} = L_{\mathrm{InfoNCE}} + L_{\mathrm{penalty}} = -\log \frac{\exp\left(\frac{(f_i \cdot f_+)}{\tau}\right)}{\sum \exp\left(\frac{(f_i \cdot f_-)}{\tau}\right)} + |f_i \cdot f_+ - p(f_+)| \quad (8)$$

$$p(f_+) = \lambda \cdot (m^2_{\mathrm{displacement}} + m^2_{\mathrm{intensity}} + m^2_{\mathrm{area\,size}}) \quad (9)$$

During the training process, guided by the target of minimizing the improved loss function, ResNet-18 and the extracted features are iteratively refined. This enables the final extracted high-level abstract features to be more representative of precipitations.

2.3 Verification Score Calculation

Since CLPFV evaluates precipitation forecasts from a holistic perspective, rather than being based on grids like existing methods (point-to-point and spatial verification methods), it means that CLPFV cannot directly use the verification scores of existing methods (e.g., TS score). Consequently, CLPFV requires a specifically designed verification score.

After stages 1 and 2, CLPFV extracts high-level feature vectors from the forecasted and observed precipitations respectively. Their similarity can ideally reflect the accuracy of precipitation forecasts: higher similarity indicates more accurate forecasts, while lower similarity suggests larger forecast errors. Therefore, the verification score of CLPFV should be directly calculated via the similarity of feature vectors.

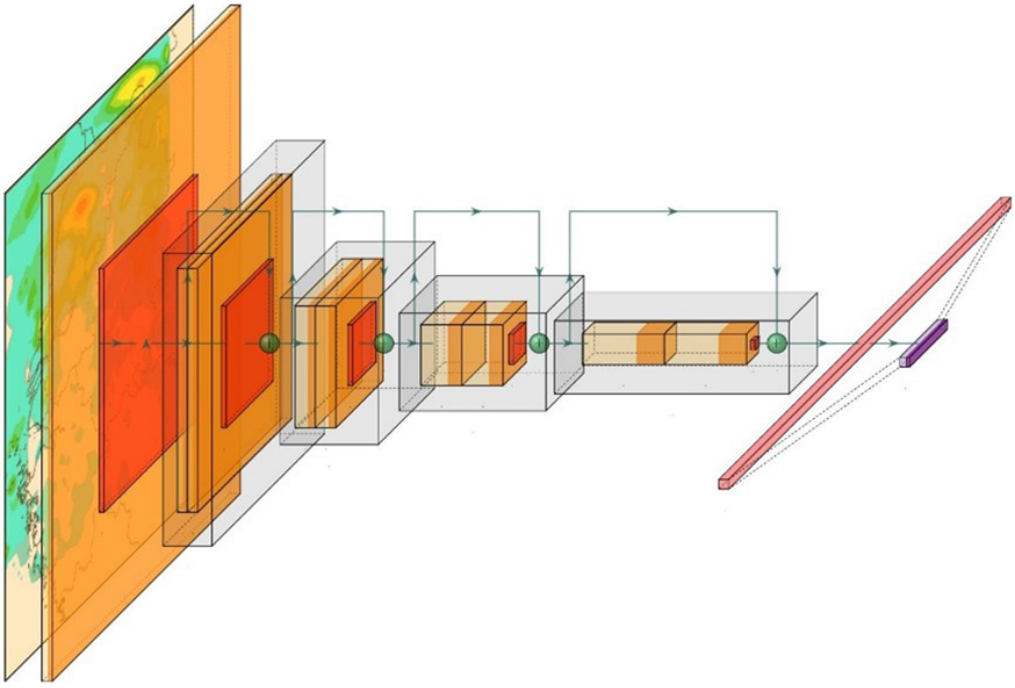


Figure 3. Structure of the network for features extraction.

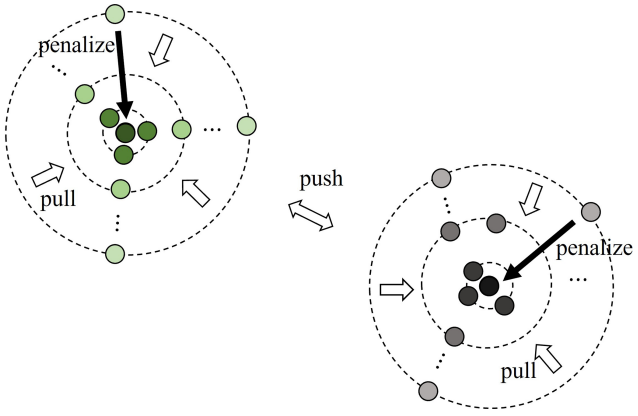


Figure 4. Schematic diagram of the improved loss function of CLPFV. Circles of the same color palette represent the data of an original precipitation and its augmented precipitations.

Considering that cosine similarity can avoid the impact of features lengths and is better to capture the essential similarity, we employ it to calculate the similarity between the forecasted and observed precipitations as CLPFV's verification score. Then Cosine similarity (verification score of CLPFV) can be expressed as Eq. (10), where f_F and f_O here are final extracted feature vectors of forecasted and observed precipitations respectively.

$$\text{similarity}\,(f_F, f_O) = \frac{f_F \cdot f_O}{|f_F||f_O|} \tag{10}$$

3 Experiments & Discussions

CLPFV is proposed to address the oversensitivity to minor forecast errors (e.g., the "double penalty" problem) while accurately reflecting varying degrees of precipitation forecast errors. Accordingly, we first designed comprehensive experiments covering diverse forecast errors scenarios, comparing CLPFV with traditional point-to-point and spatial verifications methods. In addition, to investigate the effectiveness of CLPFV in approximating expert manual verification, we conducted survey experiments with experts' verifications. The contents of experiments and discussions of results are presented by this section in detail.

3.1 Datasets

We employed the IFS (Integrated Forecasting System) dataset (Persson and Grazzini, 2007) to train CLPFV and conduct comparative experiments, covering East China region (28 to 34.5° N, 115 to 122° E) for the period 2017 to 2021, with spatial resolution of 0.125° and temporal resolution of 3 h. This dataset contains 17622 precipitations samples. All samples are divided into two parts at a ratio of around 3 : 1 through alternating sampling, with 13 500 used for training and remaining 4122 used for testing. An example of used IFS precipitation sample is given in Fig. 5.

To conduct experts survey experiments, we used the analysis and forecast products provided by CMA (China Meteorological Administration) covering the Central-South China region (22.1 to 32.6° N, 106.25 to 116.7° E) for the period

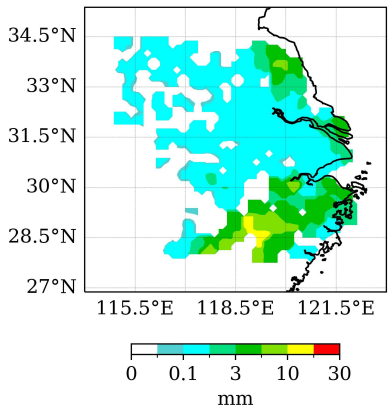


Figure 5. Examples of the IFS precipitations used in comparative experiments.

of April 2018. We specifically selected a different research area and period to conduct expert survey experiments, aiming to examine whether the contrastive learning based CLPFV possesses good generalization capability, rather thanmerely achieving high accuracy due to overfitting to the training data. The used CMA dataset contains a subset of analysis precipitation product as observed precipitations, known as CMPAS (CMA Multisource Precipitation Analysis System, Shen et al., 2018) and three subsets of forecasted precipitations from different operational numerical weather prediction models (regional short-term and mesoscale (MESO), GuangDong (GD), and ShangHai (SH)) provided by CMA. The CMA precipitation data have an identical time resolution of 3 h and were uniformly interpolated to a spatial resolution of 0.125° too. The dataset CMA consists of seven samples, each corresponding to one of the seven questions in the expert survey questionnaire. Each sample contains one observed and three forecasted precipitation products at identical timestamps. An example of the observed precipitation distribution from the CMA dataset is given in Fig. 6.

3.2 Comparative Experiments

In this research, the comparison methods for CLPFV are selected as follows.

1. Traditional point-to-point methods with POD, FAR, and TS. These verification scores are calculated based on Hits (true positives), False Alarms (false positives), and Misses (false negatives) of the precipitation forecasting confusion matrix. The POD measures how point-to-point methods accurately identify actual rainfall events, calculated as (Hits / (Hits + Misses)). The FAR measures the forecast precision and is calculated as (False Alarms / (Hits + False Alarms)). The TS provides a more balanced measure by calculating (Hits / (Hits + Misses + False Alarms)), effectively penalizing both Misses and False alarms to give a robust score of forecasting performance.

2. Spatial verification methods with Fractions Skill Score (FSS) and Structure-Amplitude-Location (SAL) metrics. FSS is a commonly used neighborhood method (Roberts and Lean, 2008), FSS calculates the verification score by comparing the ratio of precipitation grids between forecasted and observed precipitations within a specifical window. In practical applications, due to constraints such as the high cost of knowledge acquisition and limited time, FSS and other spatial verification methods often struggle to achieve full adaptation to the target task through in-depth preparations of their parameters and rules. Therefore, instead of relying on expert knowledge or adequate pre-experiments to determine the window size, we directly employed the two most used window sizes of 5 (FSS-5) and 10 (FSS-10) (Ayzel et al., 2020) in this research. As to SAL, it is a representative object-based method (Wernli et al., 2008). SAL provides the verification score by combining three indicators, i.e., combining structure (S), amplitude (A), and location (L), corresponding to the errors in precipitation area, intensity, and displacement, respectively.

CLPFV, POD, TS, and FSS have the same value range of [0, 1], whereas their higher values all indicate more accurate forecasts. But FAR, S, A, and L follow the opposite trend, i.e., their smaller absolute values represent more accurate forecasts. To facilitate experimental results presentation and comparative analysis, we normalized FAR, S, A, and L to the same value range of [0, 1], with higher normalized values now indicating more accurate precipitation forecasts too. The detailed introduction of normalization is provided in the Supplement.

To validate whether CLPFV can effectively verify precipitation forecast errors (i.e., being tolerant of minor errors and strict with significant ones), we need to establish gradient precipitation forecast error scenarios in comparative experiments to comprehensively cover different error degrees. Since the error degree in actual precipitation forecasts is difficult to foresee and control, directly obtaining uniformly distributed error gradients from real-world forecasted and observed precipitations is impractical. Therefore, we artificially generated the required series of gradient forecast error scenarios.

The specific comparative experimental setup is as follows. First, we use the original precipitations from the IFS test data as the simulated observed precipitations, and generate a series of simulated forecasted precipitations by applying positive and negative biases systematically to original precipitations. Table 1 presents the biases configurations for displacement, precipitation intensity, and area size in simulated forecasted precipitations.

The displacement bias represents the number of grids shifted in horizontal direction, e.g., +1 indicates a one-grid

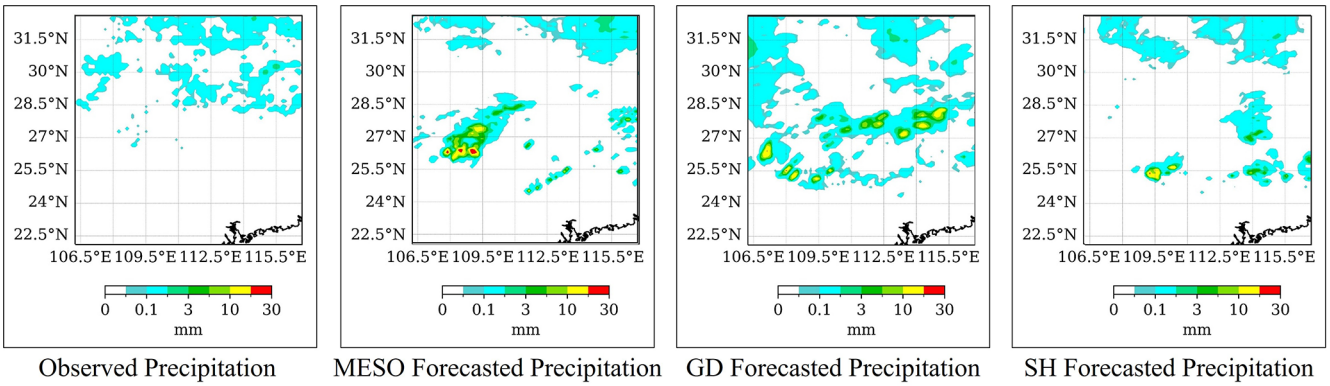


Figure 6. Examples of the CMA precipitations used in experts survey experiments.

Table 1. Applied biases for generating simulated forecasted precipitations.

Types	Biases									
Displacement	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1
Intensity	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
Area Size	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
continued	Biases									
Displacement	0	1	2	3	4	5	6	7	8	9
Intensity	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Area Size	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9

eastward shift, while −3 denotes a three-grid westward displacement shift. The intensity bias represents the multiplicative factor applied to precipitation intensity values at each grid, e.g., 0.5 reduces intensity by half, whereas 2 doubles the original values. The area size bias represents the scaling factor for the precipitation area size while maintaining a fixed centroid, e.g., 0.5 contracts the area size by 50 %, and 1.5 expands it to 150 % of the original. Figure 7 demonstrates several examples of simulated forecast precipitations with various biases configurations.

Subsequently, by inputting the simulated forecasted precipitations (i.e., generated gradient biased precipitations) and their corresponding simulated observation precipitations (i.e., the original precipitations) into verifications, we obtained the verification scores of CLPFV and comparison methods. Figure 8 presents the raw results on the IFS test data of TS, FSS-10, and CLPFV verification scores from the gradient displacement biases experiments (all raw results of comparative experiments are presented in the Supplement). According to Fig. 8, we can find that TS declined rapidly when displacement biases started to appear, reflecting its oversensitivity to minor displacement errors just as previous studies revealed. Oppositely, FSS-10 always slowly declined and thus was overly tolerant of significant displacement errors. Only CLPFV could tolerate minor errors and being strict to significant errors, thereby effectively reflecting different degrees of displacement errors.

Since directly overlaying all raw results of CLPFV and comparison methods in a single plot would make it difficult to read and analyze, we provide an optimized plot of comparative experimental results to discuss. Therefore, we conducted statistics on the raw verification scores of each verification method across different biases, calculating their average scores and standard deviations. Subsequently, by connecting the average verification scores of each bias and plotting 95 % confidence intervals, we transformed the originally dense scatters (just as Fig. 8 shows) into clear curves, making the overlay analysis of all verification methods possible. The final comparative experimental results are shown in Fig. 9. In addition, we introduced a downward-opening parabola in Fig. 9 as a benchmark reference curve for compare different PFV methods more visibly. This specific parabola was selected because its shape precisely characterizes the expected behavior of an ideal PFV method: it assigns the highest verification score in the absence of errors; permits a gradual decline for minor errors to ensure fault tolerance and avoid "double penalty" problem; and enforces an accelerated descent for larger errors to significantly penalize severe forecast failures.

According to the results of displacement biases experiments shown in Fig. 9, the verification scores of TS, POD,

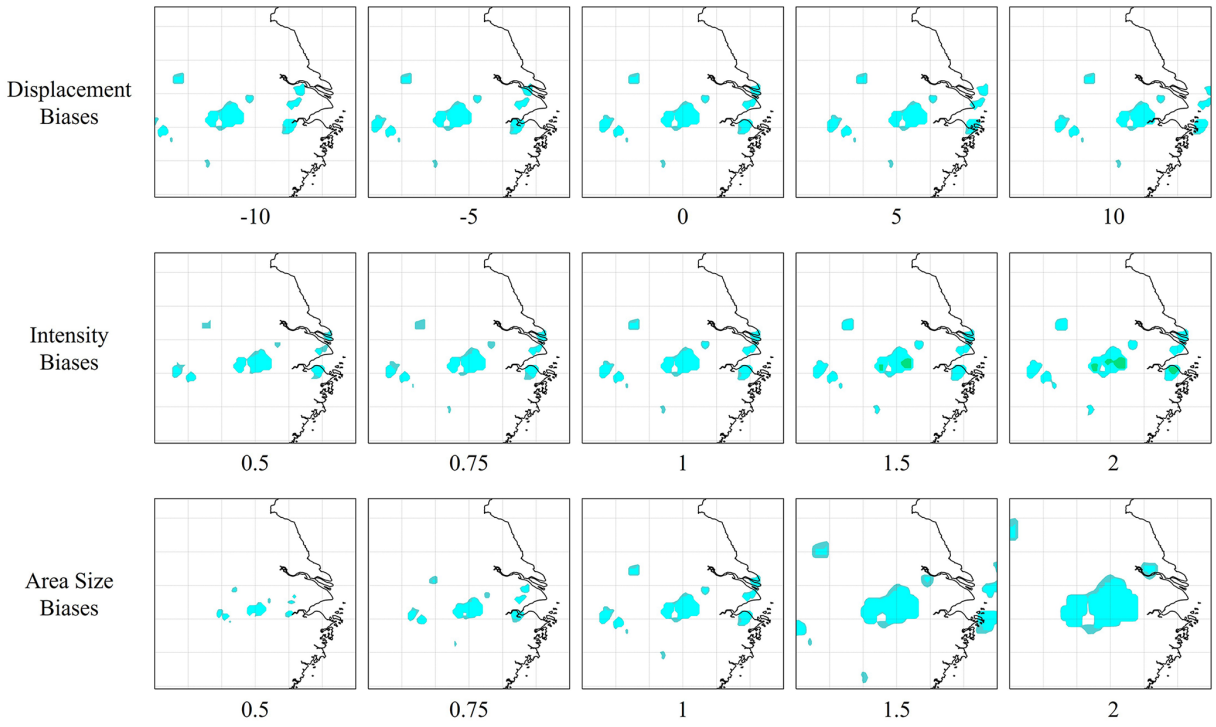


Figure 7. An example of simulated forecasted precipitations by applying gradient biases.

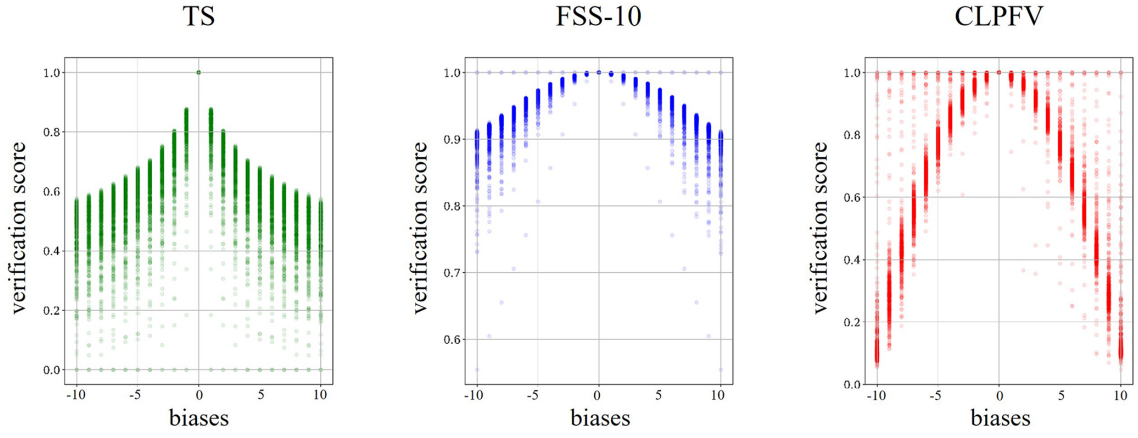


Figure 8. Raw results of gradient displacement biases experiments.

and normalized FAR dropped sharply from the centre, indicating that traditional point-to-point verification methods were overly sensitive to minor displacements. However, after reaching a certain point, their scores slow down considerably instead, suggesting insufficient penalties for substantial errors. This occurs because large displacements cause complete mismatches, stabilizing misses and false alarms. In contrast, spatial verification methods (FSS-5, FSS-10, normalized SAL) showed slow decreases, proving insensitivity to minor displacements. But they became overly tolerant of large displacements, with scores always above 0.8, overestimating forecast accuracy. The proposed CLPFV presented a rational curve and is more like reference parabola: maintaining tolerance for minor displacements while imposing stricter penalties for significant ones.

In the intensity biases experiments, both traditional point-to-point and spatial verification methods performed poorly, with their verification scores (TS, POD, normalized FAR, FSS-5, FSS-10, and normalized SAL) consistently close to 1. This is because they rely on binary thresholding, which loses intensity continuity. Specifically, even within 50 %–200 % intensity variation, most grids remained within original thresholds, preventing these methods from distinguishing intensity biased forecasts. Conversely, CLPFV successfully

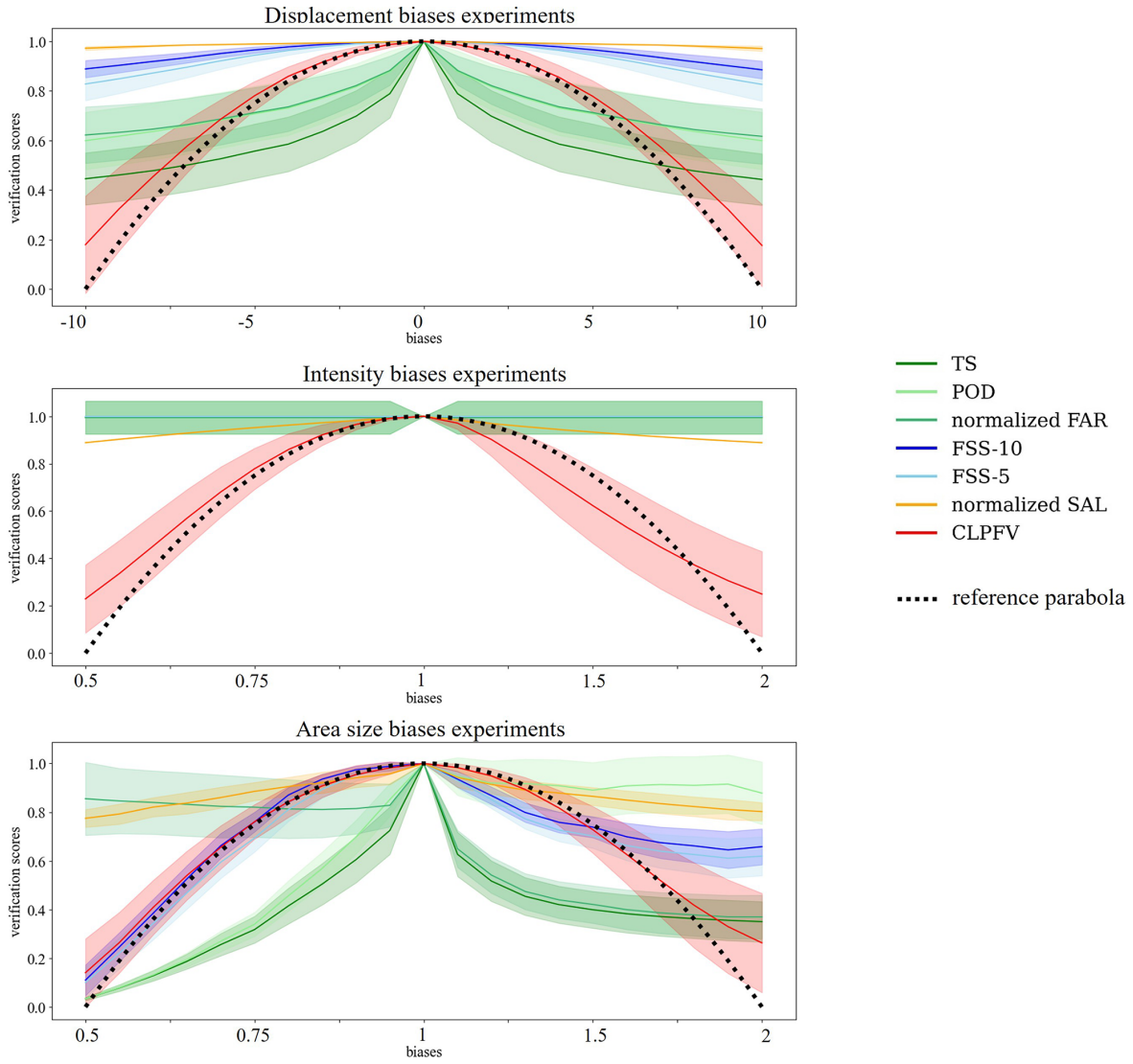


Figure 9. Results of displacement biases (top), intensity biases (middle), area size biases (bottom) experiments. The experimental results of each verification method are presented as average verification score curves with their 95 % confidence intervals.

captured gradient intensity biases, with scores systematically decreasing as bias increased, demonstrating its unique threshold-free design.

In the area size biases experiments, TS, POD, and normalized FAR remained excessively sensitive to minor biases. When areas shrank, POD and TS declined sharply due to fewer Hits and Misses, while normalized FAR stayed high because False Alarms stayed relatively constant. When areas expanded, increasing False Alarms lowered TS and normalized FAR, while POD remained stable due to unchanged Hits and Misses. As to spatial verification methods, FSS-5 and FSS-10 effectively reduced oversensitivity for area size reduction but failed to penalize expansion adequately as their verification scores decline slowed. The neighbor smoothing of FSS further inflated scores, sometimes exceeding traditional point-to-point methods like TS and normalized FAR. Normalized SAL remained above 0.9, indicating ineffectiveness against area size errors. Only CLPFV maintained robust performance, tolerating minor errors while strictly penalizing large ones and assigning appropriate scores based on bias degree.

The comparison experiments indicated that CLPFV successfully met its objectives: it showed reasonable tolerance for minor errors while strictly penalizing significant ones and reflects error degrees accurately. In contrast, traditional point-to-point methods are oversensitive, while spatial methods (with suboptimal parameters) are overtolerant of major errors. This is particularly problematic in area expansion tests, where high scores may encourage selecting models with more false alarms. Such outcomes mislead decision-

makers, weaken flood control planning, and waste disaster prevention resources.

3.3 Experts Survey Experiments

Expert manual verification is still an important benchmark for assessing the reliability of PFV methods. Therefore, we also compared CLPFV and comparison verification methods with Experts verifications in this research. To quantify how closely different verification methods align with experts, it is necessary to compare subjective expert verifications with objective verification scores under a unified standard. Accordingly, we conducted experts survey experiments as follows: First, a questionnaire is used to obtain experts' rankings of the specified precipitation forecasts. Second, to achieve the comparison of PFV's verification scores with experts manual verification, we implemented CLPFV and comparison verification methods on the same precipitation forecasts to acquire their corresponding rankings. Finally, by comparing all ranks, we could find which verification method is closer to the expert manual verification.

The questionnaire consists of seven questions, corresponding to the CMA precipitations samples mentioned in Datasets. The experts' survey was independently completed by 36 experienced forecasters. As shown in Fig. 10, experts were asked to rank the forecasted precipitations after reviewing and comparing the observed and three forecasted precipitations in a question. The ranks of each question from all experts were averaged to constitute the final expert ranking benchmark for this question. To ensure the validity of ranking, we carefully selected precipitations samples with different precipitation intensity and diverse distribution characteristics, with ensuring that the performance of each forecast model varied across different questions, avoiding tasks that were too simplistic while preventing ranking inertia. Subsequently, we applied the same verification methods (i.e., TS, POD, FAR, FSS-5, FSS-10, SAL, CLPFV). Finally, the Spearman Footrule distance (Ilyas et al., 2008) is used to quantify the similarity between each verification method and the expert ranking benchmark, and the Spearman Footrule distance SF_k is calculated with Eq. (10),

$$\mathrm{SF}_k = \frac{\sum_j^{N_Q} \left| r_{kj} - r_{\mathrm{E}j} \right|}{N_Q} \tag{11}$$

where k is the verification method, j is the question of questionnaire, N_Q is the number of questions, r_{kj} is j question's ranking of verification method k, and $r_{\mathrm{E}j}$ is j question's ranking benchmark of experts. Accordingly, a smaller Spearman Footrule distance indicates higher consistency between the verification method and experts.

The results of experts survey experiments are presented in Table 2. As shown, CLPFV had the smallest Spearman Footrule distance, demonstrating that CLPFV was significantly more similar with expert verifications compared to other methods. This confirms that CLPFV leveraged the ability of deep learning in simulating human analysis and thereby enhancing the reliability of its verifications. Furthermore, in our experts survey experiments, we specifically selected precipitation forecast verification tasks from different regions and periods. Therefore, the good performance of CLPFV also demonstrates its good generalization capability.

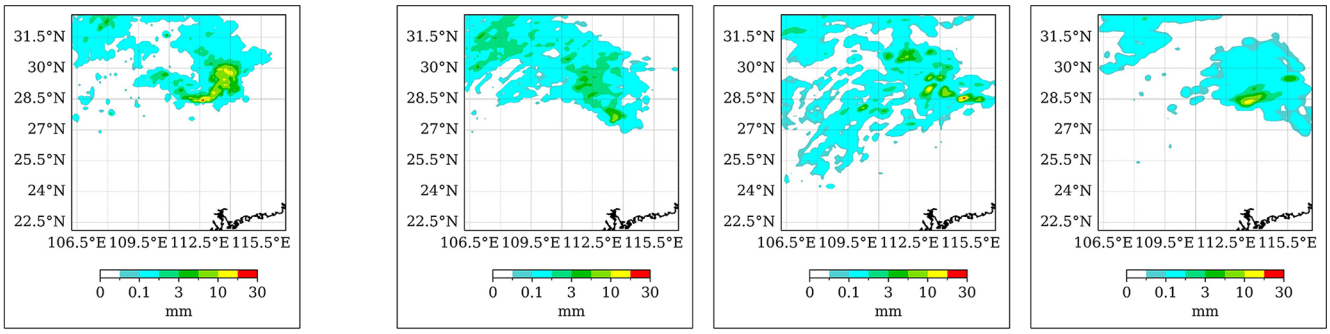
TS ranked as the second-best in experts survey experiments, indicating that TS indeed provided a more comprehensive verification than POD and FAR by simultaneously accounting for Hits, Misses, and False Alarms. Moreover, the good performance of TS also suggested that traditional point-to-point method can still yield reasonable verifications, despite some situations like "double penalty" problem.

In contrast, spatial verification methods underperformed compared to traditional point-to-point verification methods. This result seems to contradict common sense, but in fact, it is not inconsistent with existing studies. In the studies on spatial verification methods, these methods and their parameters are often specifically optimized for experimental cases, thereby appearing their upper limits of verification capabilities. However, in practical applications, spatial verification methods are highly likely to lack the required expert knowledge and pre-experimentation, resulting in mismatched parameters and rules that undermine the verification accuracy. Therefore, previous studies have mostly demonstrated the excellent performance of spatial verification methods under ideal conditions, Therefore, previous studies have mostly demonstrated the excellent performance of spatial verification methods under ideal conditions, while our research highlights their limitations in practical applications: without sufficient optimization, the performance of spatial verification methods may be inferior to that of traditional point-to-point methods. This also manifests that the generalization capability and practicality of spatial verification methods still need to be improved.

To illustrate experimental results more intuitively, we selected Question 2 of questionnaire as a typical case to analyze the differences between the rankings of PFV methods and the experts ranking benchmark.

Figure 11 shows the experts' rankings for the three numerical prediction models (MESO, GD, and SH) in Question 2. As seen in Fig. 11, the observed precipitation has two key distribution characteristics: a strip-shaped region of heavy precipitation in the centre of the study area and light precipitation across the entire northwest regions. As to their forecasts, GD best captured the precipitation distribution, particularly for the heavy precipitation region. MESO overestimated the heavy precipitation, while SH underestimated both the precipitation coverage and intensity. Thus, the expert ranking of them is GD (1st) > MESO (2nd) $\gg$ SH (3rd).

Table 3 shows the performance of each PFV method for Question 2. Due to their dependence on threshold settings, traditional point-to-point methods (TS, POD, FAR) failed to reflect the distribution of heavy precipitation. As a result,



Please rank the three forecasted precipitations in descending order of accuracy: (1)____ (2)____ (3)____

Figure 10. A question example of expert ranking questionnaire.

Table 2. Results of experts survey experiments.

PFV methods	POD	FAR	TS	FSS-5	FSS-10	SAL	CLPFV
Spearman Footrule distance	2.00	2.00	1.14	2.00	2.29	1.43	0.57

they all incorrectly ranked MESO as the best forecast. Notably, FAR, which only considers Hits and False Alarms, even incorrectly ranked SH as the second-best forecast. As to spatial verification methods, neighborhood-based methods (FSS-5 and FSS-10) produced the same erroneous rankings. Due to their spatial smoothing operations, the underestimation of precipitation in SH and the overestimation of heavy precipitation in MESO were both mitigated, leading to their incorrect ranking of SH and MESO. Although object-based method (SAL) was slightly better than FSS-5 and FSS-10, it still incorrectly ranked MESO as the best since its focus on overall precipitation features and ignorance of inner heavy precipitation distributions. In contrast, only CLPFV fully matched the expert ranking benchmark, especially its low score for SH accurately reflected the fact that SH's forecast was significantly worse. This case analysis demonstrates that CLPFV effectively captured precipitation distribution characteristics, yielding more reliable verification results.

4 Conclusions

To address the issues of point-to-point PFV methods being overly sensitive to minor errors and the issues of spatial PFV methods requiring appropriate parameters and rules with limiting their practicality, as well as tackling the challenge of lacking labeled samples for training deep neural networks, we propose CLPFV by introducing self-supervised contrastive learning into precipitation forecast verification. First, CLPFV constructs positive and negative sample pairs by precipitations' multi augmentations of displacement, intensity, and area. Then, it trains the ResNet-18 network using an improved contrastive loss function to extract high-level abstract features of the forecasted and observed precipitations respectively. Finally, the verification score of CLPFV is calculated through the feature-based cosine similarity.

We compared CLPFV with traditional point-to-point and spatial verification methods with comprehensive gradient biases experiments. The results show that CLPFV could effectively assign the appropriate verification scores to all degrees of precipitation forecast errors. Furthermore, CLPFV had the highest consistency in the ranking results of experts' survey, confirming that CLPFV matched expert manual verification better.

Notably, CLPFV also serves as a feasible framework when evaluating the prediction of other meteorological variables or environmental phenomena. Depending on the specific requirements of forecasting or prediction tasks, e.g., $PM_{2.5}$ forecast and soil mapping, the elements of CLPFV (such as ResNet-18 deep learning model, InfoNCE loss function, and Cosine similarity calculation) can be changed to adapt with various architectures. This flexibility underscores CLPFV's methodological contribution as a generalizable paradigm of forecast verification in the broad earth science field.

However, we should also note that CLPFV still has potential shortcomings. For example, as a "black-box" method based on deep learning, CLPFV cannot explicitly reveal the specific considerations and lacks interpretability, while clarifying specific issues within forecast models is one of the important goals of precipitation forecast verification. Additionally, quantitative precipitation forecasting is usually based on continuous time periods in current practical applications, but current point-to-point and spatial verification methods heavily rely on subjective judgment and are difficult to quanti-

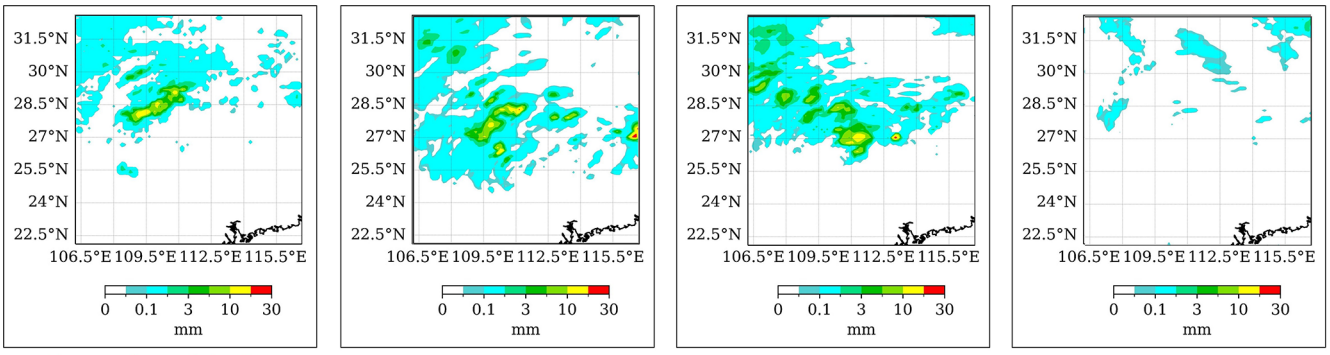


Figure 11. The expert ranking benchmark and precipitations of Question 2.

Table 3. Verification performances of PFV methods for Question 2 in experts survey experiments.

	GD verification score	MESO verification score	SH verification score	GD rank	MESO rank	SH rank
TS	0.28	0.38	0.22	2	1	3
POD	0.75	0.79	0.37	2	1	3
Normalized FAR	0.31	0.42	0.37	3	1	2
FSS-5	0.59	0.70	0.63	3	1	2
FSS-10	0.65	0.77	0.76	3	1	2
Normalized SAL	0.75	0.80	0.55	2	1	3
CLPFV	0.77	0.53	0.10	1	2	3

tatively evaluate forecasts across multiple consecutive time steps. Therefore, it is one of the important future research directions of CLPFV that developing it to enable quantitative verifications across continuous time periods. Furthermore, for more interdisciplinary fields of artificial intelligence and Earth sciences, such as ecological modelling, soil mapping, disaster prediction and et al., CLPFV could provide a valuable refence for improving verification methods in these fields and may even pave the way for developing a general verification method applicable across the broader Earth sciences domain. This also represents an important direction for future research.

Code and data availability. The source code and data of this work have been packaged and can be found at Zenodo platform https://doi.org/10.5281/zenodo.16777790 (NUDT, 2025) The ReadMe file can be also found at https://doi.org/10.5281/zenodo.16777790 (NUDT, 2025), which introduces the functions of each code file and the data required for the experiments.

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/gmd-19-6027-2026-supplement.

Author contributions. Y.W., S.H., and Q.L. co-conceived the research, and performed model development. Y.W. and S.H. performed simulations, analysis, and wrote original manuscript. Y.W. revised manuscript. X.P. helped with model improvements and H.C. and K.Z. helped analysis. L.W. and S.L. helped with manuscript and figures. Q.L. obtained funding. All authors contributed to results interpretation, manuscript writing, and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Financial support. This research was funded by the National Natural Science Foundation of China (grant nos. 42075139, U2242201, and 41305138), the China Postdoctoral Science Foundation (grant

no. 2017M621700), Hunan Province Natural Science Foundation (grant nos. 2021JC0009 and 2021JJ30773).

Review statement. This paper was edited by Rohitash Chandra and reviewed by two anonymous referees.

References

Ayzel, G., Scheffer, T., and Heistermann, M.: RainNet v1.0: a convolutional neural network for radar-based precipitation nowcasting, Geosci. Model Dev., 13, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020, 2020.

Cassola, F., Ferrari, F., and Mazzino, A.: Numerical simulations of Mediterranean heavy precipitation events with the WRF model: a verification exercise using different approaches, Atmos. Res., 164, 210–225, https://doi.org/10.1016/j.atmosres.2015.05.010, 2015.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.: A simple framework for contrastive learning of visual representations, in: Proceedings of the 8th International Conference on Learning Representations (ICLR), 26 April–1 May 2020, Addis Ababa, Ethiopia, 1597–1607, arXiv, https://doi.org/10.48550/arXiv.2002.05709 2020.

Chen, Y., Wang, Y., Huang, G., and Tian, Q.: Coupling physical factors for precipitation forecast in China with graph neural network, Geophys. Res. Lett., 51, e2023GL106676, https://doi.org/10.1029/2023GL106676, 2024.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts, Mon. Weather Rev., 134, 1772–1784, https://doi.org/10.1175/MWR3145.1, 2006.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J.: The setup of the MesoVICT project, B. Am. Meteorol. Soc., 99, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1, 2018.

Dorninger, M., Ghelli, A., and Lerch, S.: Recent developments and application examples on forecast verification, Meteorol. Appl., 27, e1934, https://doi.org/10.1002/met.1934, 2020.

Ebert, E. E. and Gallus Jr., W. A.: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification, Weather Forecast., 24, 1401–1414, https://doi.org/10.1175/2009WAF2222252.1, 2009.

Gidaris, S., Singh, P., and Komodakis, N.: Unsupervised representation learning by predicting image rotations, in: Proceedings of the 6th International Conference on Learning Representations (ICLR), 30 April–3 May 2018, Vancouver, Canada, arXiv, https://doi.org/10.48550/arXiv.1803.07728, 2018.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of spatial forecast verification methods, Weather Forecast., 24, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1, 2009.

Gilleland, E., Ahijevych, D., Brown, B. G., and Ebert, E. E.: Verifying forecasts spatially, B. Am. Meteorol. Soc., 91, 1365–1376, https://doi.org/10.1175/2010BAMS2819.1, 2010.

Gofa, F., Boucouvala, D., Louka, P., and Flocas, H.: Spatial verification approaches as a tool to evaluate the performance of high resolution precipitation forecasts, Atmos. Res., 208, 78–87, https://doi.org/10.1016/j.atmosres.2017.09.021, 2018.

Gofa, F., Flocas, H., Louka, P., and Samos, I.: A coherent approach to evaluating precipitation forecasts over complex terrain, Atmosphere, 13, 1164, https://doi.org/10.3390/atmos13081164, 2022.

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M.: Bootstrap your own latent – a new approach to self-supervised learning, in: Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 6–12 December 2020, online, 21271–21284, 2020.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 27–30 June 2016, Las Vegas, USA, 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016.

Ilyas, I. F., Beskales, G., and Soliman, M. A.: A survey of top-k query processing techniques in relational database systems, ACM Comput. Surv., 40, 1–58, https://doi.org/10.1145/1391729.1391730, 2008.

Jain, S., Scaife, A. A., Shepherd, T. G., Deser, C., Dunstone, N., Schmidt, G. A., Trenberth, K. E., and Turkington, T.: Importance of internal variability for climate model assessment, npj Clim. Atmos. Sci., 6, 68, https://doi.org/10.1038/s41612-023-00389-0, 2023.

Lee, S.-H., Kim, S.-W., Angevine, W. M., Bianco, L., McKeen, S. A., Senff, C. J., Trainer, M., Tucker, S. C., and Zamora, R. J.: Evaluation of urban surface parameterizations in the WRF model using measurements during the Texas Air Quality Study 2006 field campaign, Atmos. Chem. Phys., 11, 2127–2143, https://doi.org/10.5194/acp-11-2127-2011, 2011.

Li, L., Shao, A., and Qiu, X.: Short-term forecast large-scale error characteristics and their relationship with precipitation forecast skill under two rainfall regimes, Atmos. Res., 298, 107152, https://doi.org/10.1016/j.atmosres.2023.107152, 2024.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J.: Self-supervised learning: generative or contrastive, IEEE T. Knowl. Data En., 35, 857–876, https://doi.org/10.1109/TKDE.2021.3090866, 2021.

Mittermaier, M. P. and Roberts, N. M.: Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score, Weather Forecast., 25, 343–354, https://doi.org/10.1175/2009WAF2222260.1, 2010.

NUDT: CLPFV, Zenodo [code, data set], https://doi.org/10.5281/zenodo.16777790, 2025.

Oord, A., Li, Y., and Vinyals, O.: Representation learning with contrastive predictive coding, arXiv [preprint], https://doi.org/10.48550/arXiv.1807.03748, 11 July 2018.

Persson, A. and Grazzini, F.: User guide to ECMWF forecast products, ECMWF Meteorological Bulletins M3.2, ECMWF, Reading, UK, 115 pp., https://doi.org/10.21957/m1cs7h, 2007.

Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Mon. Weather Rev., 136, 78–97, https://doi.org/10.1175/2007MWR2123.1, 2008.

Rossa, A., Nurmi, P., and Ebert, E.: Overview of methods for the verification of quantitative precipitation forecasts, in: Precipitation: Advances in Measurement, Estimation and Prediction, edited by: Michaelides, S., Springer, Berlin, Heidelberg, Ger-

many, 419–452, https://doi.org/10.1007/978-3-540-77655-0_16, 2008.

Saavedra Valeriano, O. C., Koike, T., Yang, K., Graf, T., Li, X., Wang, L., and Han, X.: Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts, Water Resour. Res., 46, W10544, https://doi.org/10.1029/2010WR009502, 2010.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H.: A theoretical analysis of contrastive unsupervised representation learning, in: Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019, arXiv [preprint], https://arxiv.org/abs/1902.09229, 2019.

Shen, Y., Hong, Z., Pan, Y., Yu, J., and Maguire, L.: China's 1 km merged gauge, radar and satellite experimental precipitation dataset, Remote Sens., 10, 264, https://doi.org/10.3390/rs10020264, 2018.

Qiu, Y., Feng, J., Zhang, Z., Zhao, X., Li, Z., Ma, Z., Liu, R., and Zhu, J.: Regional aerosol forecasts based on deep learning and numerical weather prediction, npj Clim. Atmos. Sci., 6, 71, https://doi.org/10.1038/s41612-023-00397-0, 2023.

van der Plas, E., Schmeits, M., Hooijman, N., and Kok, K.: A comparative verification of high-resolution precipitation forecasts using model output statistics, Mon. Weather Rev., 145, 4037–4054, https://doi.org/10.1175/MWR-D-16-0256.1, 2017.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL – a novel quality measure for the verification of quantitative precipitation forecasts, Mon. Weather Rev., 136, 4470–4487, https://doi.org/10.1175/2008MWR2415.1, 2008.

Wernli, H., Hofmann, C., and Zimmer, M.: Spatial forecast verification methods intercomparison project: application of the SAL technique, Weather Forecast., 24, 1472–1484, https://doi.org/10.1175/2010BAMS2819.1, 2009.

Xu, L., Chen, N., Chen, Z., Zhang, C., and Yu, H.: Spatiotemporal forecasting in Earth system science: methods, uncertainties, predictability and future directions, Earth-Sci. Rev., 222, 103828, https://doi.org/10.1016/j.earscirev.2021.103828, 2021.

Xu, X., Sun, X., Han, W., Zhong, X., Chen, L., Gao, Z., and Li, H.: FuXi-DA: a generalized deep learning data assimilation framework for assimilating satellite observations, npj Clim. Atmos. Sci., 8, 156, https://doi.org/10.1038/s41612-025-01039-3, 2025.

Zhu, K., Zhang, C., Xue, M., and Yang, N.: Predictability and skill of convection-permitting ensemble forecast systems in predicting the record-breaking "$21 \cdot 7$" extreme rainfall event in Henan Province, China, Sci. China Earth Sci., 65, 1879–1902, https://doi.org/10.1007/s11430-022-9961-7, 2022.