Comparative Evaluation of ecPoint and Local EMOS for CMA-GEPS Precipitation Forecast over Eastern China

Sonum Stejik[1,2], Pu Liu[3], Wang Jialing[6] and Wang Yong[1,4,5]

[1]*School of Atmospheric Sciences, Nanjing University of Information Science and Technology, Nanjing, China*
[2] *Meteorological Bureau of Xigaze City, Xizang, China*
[3] *Nanjing Meteorological Bureau, Nanjing, China*
[4] *Earth System Modeling and Prediction Centre, China Meteorological Administration, Beijing, China*
[5] *State Key Laboratory of Severe Weather Meteorological Science and Technology, Beijing, China*
[6] *Academic Affairs Office, Nanjing University of Information Science and Technology, Nanjing, China*

1. Supplement Method

1.1 ecPoint

ecPoint is a novel point-based innovative statistical post-processing technique primarily proposed by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Pillosu & Hewson, 2017; Hewson & Pillosu, 2021). The ERA5-ecPoint reanalysis product, based on ECMWF's ERA5 data, is the first-ever point-scale probabilistic global reanalysis product (Hersbach et al., 2020).

ecPoint is a non-parametric post-processing approach that relies on the forecast error rate (FER), calculated from the difference between observed precipitation r_0 at test sites and the corresponding grid forecast G. The method applies a decision tree to assign grid points to different statistical mapping functions:

$$\mathrm{FER} = (r_0 - G)/G$$

Each node in the decision tree represents a grid cell, storing statistical relationships derived from historical forecast and observation data. For node j, the mapping function Mj is defined based on the FER distribution:

$$C_j = 1 + \int_{-1}^{\infty} x\mathbf{M}_j \mathrm{d}x$$

where x represents FER and Mj is the probability density function (PDF) for the j-th node. The raw numerical forecast for each ensemble member is corrected by incorporating the expected grid-scale bias correction factor (i.e., the sub grid variance of the sum of grid-scale NWP biases). For the i-th ensemble member, C_j represents the bias correction factor, C_j denotes the raw precipitation forecast value, and FER is transformed into $\mathrm{M}_i(\mathrm{FER})$ representing the mapping function selected for that member. The point-specific precipitation forecasts Fi for ensemble member i are derived as:

$$\mathrm{F}_i(r) = \big(1 + \mathrm{M}_i(\mathrm{FER})\big)\mathrm{G}_i \quad \mathrm{M}_i \in \{\mathrm{M}_1, \dots, \mathrm{M}_m\}$$

The final probabilistic precipitation forecast vector $\mathbf{F}$ for a given grid location is obtained by aggregating all n ensemble members:

$$\mathbf{F} = \frac{1}{n}\left(\sum_{i=1}^{n} \mathbf{F}_i(r)\right)$$

1.2 EMOS

The Ensemble Model Output Statistics (EMOS) method is an extension of the traditional Model Output Statistics (MOS) approach for quantitative forecasting. Building on MOS, EMOS provides a parameterized post-processing framework that links ensemble forecasts to a chosen predictive probability distribution. The method first specifies an appropriate probability distribution for the forecast variable and then uses a link function to relate the distribution parameters to the ensemble forecasts. The parameters are estimated by an optimization procedure and used for correcting forecast biases (Gneiting et al., 2005; Baran et al., 2016).

For continuous variables such as temperature or wind speed, EMOS typically assumes a normal or truncated normal distribution. However, for precipitation—which is non-negative, highly skewed, and has a point mass at zero—the normal distribution is unsuitable. Instead, EMOS adopts a left-censored, shifted gamma (CSG) distribution to model precipitation. The CSG distribution accommodates continuous values that can be positive or zero, with left-censoring at zero (Scheuerer, 2015; Scheuerer & Hamill, 2015). This formulation underpins the CSG-based EMOS model proposed by Baran and Nemoda (2016).

Let $P_{\kappa,\theta}$ denote the cumulative distribution function (CDF) of the gamma distribution Γ(κ, θ) with shape parameter κ > 0 and scale parameter θ > 0. Its probability density function (PDF) is defined as:

$$P_{\kappa,\theta}(x) := \begin{cases} \frac{x^{\kappa-1}e^{-\frac{x}{\theta}}}{\theta^{\kappa}\Gamma(\kappa)}, & x > 0, \\ 0, & \text{otherwise}, \end{cases}$$

Here, Γ(κ) denotes the value of the gamma function at κ. The parameters κ and θ correspond one-to-one with the mean μ > 0 and standard deviation σ > 0 of the gamma distribution:

$$\kappa = \frac{\mu^2}{\sigma^2} \quad \theta = \frac{\sigma^2}{\mu},$$

By introducing a shift parameter δ > 0, the support range of gamma distribution can be extended to negative values, and with left-censored at zero yields the CSG distribution Γ_0 (κ, θ, δ). Its CDF and generalized PDF are defined as:

$$C^0_{\kappa,\theta,\delta}(x) = \begin{cases} C_{\kappa,\theta}(x+\delta), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

$$\mathrm{P}_{k,\theta,\delta}(x) = \begin{cases} [1 - \mathrm{C}_{k,\theta}(\delta)]\mathrm{P}_{k,\theta}(x+\delta), & x > 0 \\ \mathrm{C}_{k,\theta}(\delta), & x = 0, \\ 0, & x < 0 \end{cases}$$

Here, κ, θ, and δ are the shape, scale, and shift parameters, respectively, and $C_{\kappa,\theta}$ denotes the CDF of the original gamma distribution Γ (κ, θ). For a quantile x_p(0 < p <1), when $p \leq C_{k,\theta}(\delta)$, indicating zero probability of precipitation, $x_p = 0$; while when $p > C_{k,\theta}(\delta)$, indicating precipitation occurances, x_p is the solution to the equation $p = C_{k,\theta}(x_p + \delta)$.

In the EMOS_CSG model, the mean μ and variance σ² of the underlying gamma distribution are linked to ensemble forecasts as:

$$\mu = a_0 + a_1 f_1 + a_2 f_2 \cdots + a_{\mathrm{M}} f_{\mathrm{M}} \quad \text{and} \quad \sigma^2 = b_0 + b_1 \bar{f},$$

Where M is the number of ensemble members, f_M represents the cumulative precipitation forecast for ensemble member m at a specific location, time, and lead time, and $\bar{f}$ denotes the ensemble mean.

1.3 Verification Methods

The Continuous Ranked Probability Score (CRPS) is one of the most widely used metrics for evaluating ensemble and probabilistic forecasts in atmospheric science. It jointly assesses calibration (the statistical consistency between forecasts and observations) and sharpness (the concentration of the predictive distribution). In many statistical post-processing frameworks, forecasts are expressed as ensemble samples or mixture distributions (e.g., zero-inflated distributions for precipitation), which often lack closed-form PDFs. In contrast, the CDF provides a universal basis for comparison (Wilks, 2011, Section 8.5.1). A key advantage of CRPS is that it evaluates forecasts directly in terms of their CDFs, thereby accounting for and penalizing the trade-off between sharpness and calibration. Lower CRPS values indicate better predictive skill. Mathematically, CRPS is defined as:

$$\mathrm{CRPS}(C, y) := \int_{-\infty}^{\infty} \left(C(x) - 1_{\{y \geq x\}}\right)^2 \mathrm{d}y$$

where $C(x)$ denotes the predicted CDF, y is the observed value, and $1_{\{y \geq x\}}$ is the indicator function for the observation.

The Continuous Ranked Probability Skill Score (CRPSS) quantifies the relative skill improvement of a predictive system or model compared to a reference benchmark. Higher CRPSS values indicate superior performance relative to the benchmark:

$$\mathrm{CRPSS} = 1 - \frac{CRPS_{forcast}}{CRPS_{reference}}$$

The Brier Score (BS) evaluates the accuracy of probabilistic predictions for binary events and is one of the most commonly used scalar performance metrics in meteorology (Wilks, 2011, Section 8.4.2). It computes the mean squared error between the predicted probability C(t) and the binary observation y at a given threshold t. BS ranges from 0 to 1, where 0 indicates a perfect forecast and 1 denotes the worst performance, and it is defined as:

$$\mathrm{BS}(C, y; t) = \left(C(t) - 1_{\{y \geq t\}}\right)^2$$

The Relative Operating Characteristic (ROC) curve, first introduced by Mason (1982) in meteorological applications, assesses the discrimination ability of probabilistic or binary classification forecasts. It illustrates how the true positive rate (TPR) varies with the false positive rate (FPR) across probability thresholds. The Area Under the Curve (AUC) provides a scalar measure of performance: values close to 1 indicate excellent discriminatory skill; a value of 0.5 corresponds to no skill; and values below 0.5 indicate performance worse than random chance.

Reliability diagrams provide a graphical assessment of forecast calibration. The horizontal axis represents forecast probability, and the vertical axis shows the observed event frequency. A perfectly calibrated forecast lies on the diagonal. Curves above the diagonal indicate under confidence (observations exceed predicted probabilities), while curves below indicate overconfidence (predictions exceed observed frequencies).

The sharpness diagram complements the reliability diagram by illustrating the distribution of issued forecast probabilities. The horizontal axis shows the forecast probability, and the vertical axis displays its relative frequency. High sharpness is indicated by a U-shaped distribution concentrated near 0 or 1, reflecting strong forecast certainty. Conversely, probabilities clustered around 0.5 indicate low sharpness and high uncertainty.

For the verification method, we use the Continuous Ranked Probability Score (CRPS) and the Ranked Probability Skill Score (CRPSS) to evaluate the overall probabilistic skill for continuous variables. The Brier Score (BS) and the receiver operating characteristic (ROC) curve, quantified by the area under the curve (AUC), are applied to assess the reliability and discrimination of event-based probabilistic forecasts. In addition, reliability diagrams and sharpness histograms are used to examine the calibration and spread of the probabilistic and ensemble forecasts. Additionally, we display both the ensemble-averaged RMSE and spread to provide an alternative means of verifying whether the reliability of the two methods. To verify whether the predictive capability aligns with AUC or ROC values, the TS score will be employed.

2. Supplement Discussion

The novelty of this study lies in demonstrating that ecPoint outperforms even a locally optimized EMOS. While Hemri et al. (2022) showed that ecPoint is superior to a global EMOS configuration—raising the possibility that the benchmark was overly general—our results indicate that ecPoint also exceeds the performance of Local EMOS. This constitutes a stronger result, suggesting that the advantage of ecPoint is not merely attributable to spatial averaging in EMOS, but to its fundamentally different treatment of non-Gaussian precipitation behaviour.

The BS score's sensitivity to sample imbalance further complicates interpretation. For rare precipitation events (e.g., thresholds >10 mm per 12 h), the score tends to be extremely low, which may explain unusual patterns across thresholds and lead times. ecPoint additionally removes forecasts below 1 mm to mitigate artefacts from observational discretisation (as recommended in Trotta, 2024; Hewson & Pillosu, 2021). This practice, while improving the statistical properties of FER distributions, reduces light-rainfall representation and contributes to differences in forecast spread relative to EMOS. ecPoint also extends precipitation forecasts beyond the original GEPS pattern, but more conservatively. Its spatial patterns appear smoother when interpolated to stations, and because station density is high, total precipitation sums across stations may exceed those from EMOS for the same grid cell. This leads to reduced high-threshold accuracy for certain lead times despite ecPoint's overall stronger reliability. Two case studies of convective events demonstrate that EMOS can forecast scattered or isolated light rainfall outside the GEPS-predicted region. This capability arises from EMOS's treatment of missing values, which effectively fills observational gaps and broadens forecast coverage. However, this often results in false alarms for low-threshold events, which explains its weaker BS performance at low thresholds.

Training and validation experiments show that ecPoint is highly efficient despite incorporating more

forecast factors. It is computationally lighter and faster than EMOS, offering advantages for real-time operational applications. ecPoint's strong resolution at multiple thresholds makes it particularly suitable for extreme-precipitation identification and early-warning applications. Once corrected with ecPoint, CMA-GEPS forecasts for Eastern China provide significantly enhanced early-warning capability for hazardous precipitation. As highlighted by the model developers, this improvement supports disaster mitigation efforts, particularly for flash-flood risk reduction and community preparedness.

By combining the verification results of various probabilistic forecasts with analyses of several representative weather cases, we find that both post-processing methods not only provide systematic improvements for large-scale precipitation, but also substantially enhance the skill of convective-scale precipitation forecasts. Further work is needed to couple these two post-processing techniques with convective-scale ensemble prediction systems and to systematically evaluate their performance across different weather regimes.

CRPS, BS, RMSE, and Spread values exhibit 12-hour periodic oscillations, with scoring performance reaching its peak at 12h, 36h, 60h, and 84h. Firstly, the base model itself oscillates due to climate-driven diurnal variations in daytime and nighttime precipitation forecast accuracy. Consequently, the outputs from both post-processing methods also exhibit varying degrees of fluctuation relative to the base model scores. Second, since the ecPoint post-processing method exclusively utilizes numerical model data from the 12-hour forecast timeframe (corresponding to nighttime in Beijing time) as training data, it achieves higher forecast accuracy for nighttime compared to daytime. This likely explains why ecPoint exhibits more pronounced oscillations in RMSE and Spread values.

3.Supplement Figures and explanation

Localized Torrential Rainfall in Jiangsu–Anhui (7 July 2023)

Figure S3 shows precipitation forecasts at multiple percentiles. GEPS exhibits limited percentile differentiation—reflecting low ensemble dispersion. Post-processing substantially increases spread. ecPoint and EMOS both improve precipitation occurrence forecasts; EMOS accurately identifies scattered light rainfall while ecPoint broadens high-intensity regions, causing some false alarms in Shandong, Jiangxi, and Zhejiang. EMOS performs better at capturing station-level heavy precipitation, whereas ecPoint provides smoother spatial patterns but risks over-prediction.

Both post-processing methods improve Threat Scores across thresholds. ecPoint yields higher TS for light, moderate, and heavy precipitation, whereas EMOS performs better at thresholds above 50.0 mm. For the 75th-percentile ensemble (Figure S5), ecPoint delivers the highest TS at most thresholds except 100.0 and 140.0 mm, consistent with its AUC performance. EMOS excels at thresholds above 100.0 mm, reflecting its stronger discrimination capability for extreme events.

Figure S6 compares ensemble spread and RMSE. All three systems exhibit similar fluctuation patterns, but GEPS suffers from severely underestimated spread. Post-processing effectively inflates the spread, correcting under-dispersion. ecPoint demonstrates the tightest alignment between spread and RMSE, indicating superior representation of uncertainty. GEPS shows increasing spread and RMSE with lead time, whereas ecPoint and EMOS, though fluctuating, do not show systematic error growth. ecPoint additionally displays clear diurnal periodicity in 12-hour precipitation.

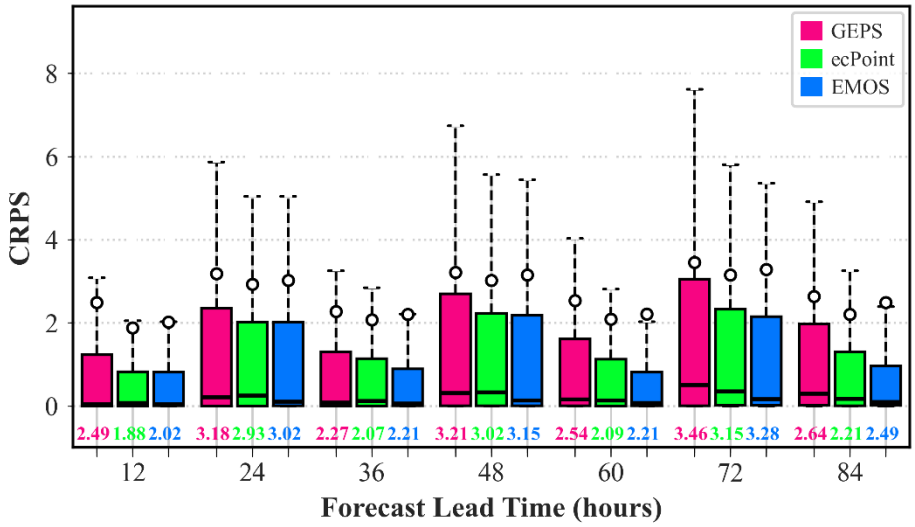

Fig. S1 Box plot of CRPS values for different forecast lead times. White circles represent the mean values, thick black lines represent the median values, and the bottom section shows the CRPS mean values for each method across different forecast lead times (mm/12h).

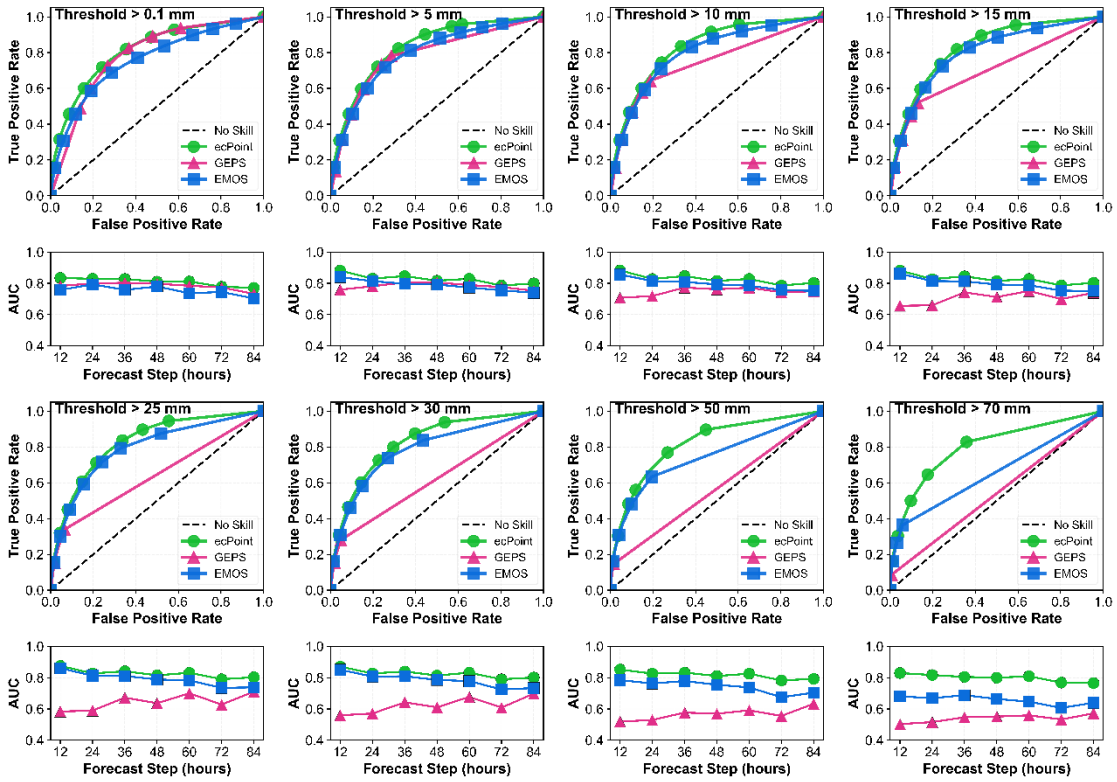

Fig. S2 ROC Curves for different thresholds and AUC values under more varied thresholds and

forecast leadtimes: two post-processing approaches and the raw model (mm/12h).

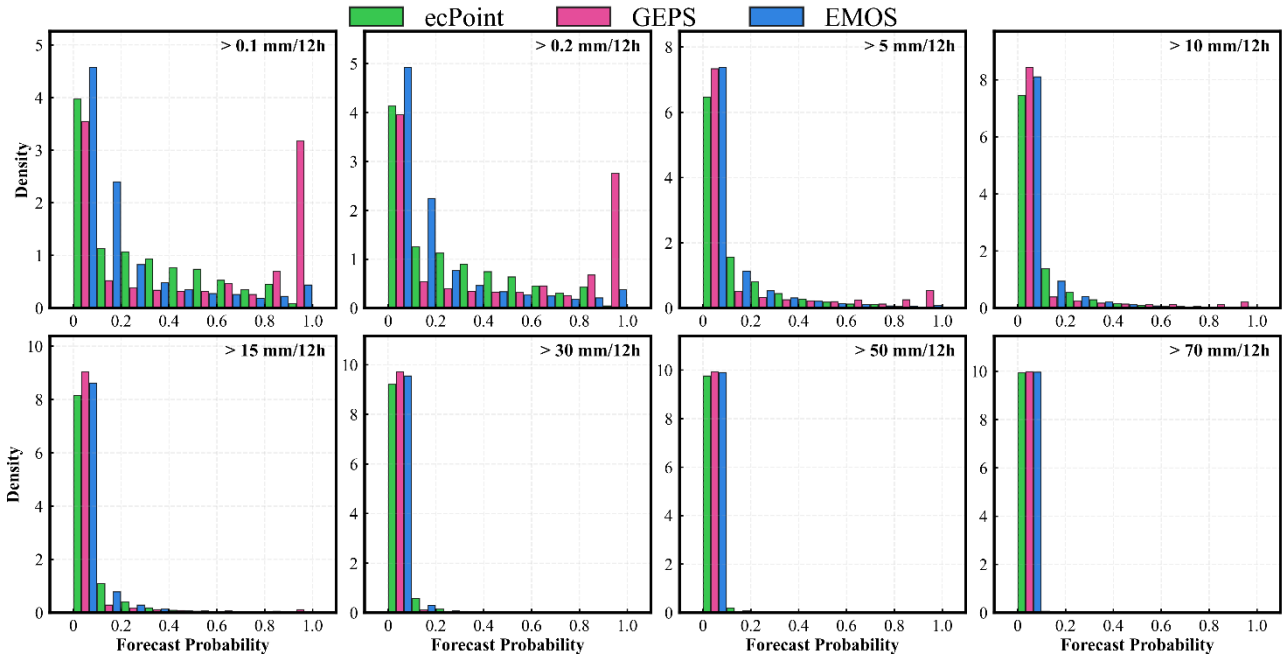

Fig. S3 Sharpness profiles of the raw model and two post-processing approaches at more thresholds (mm/12h).

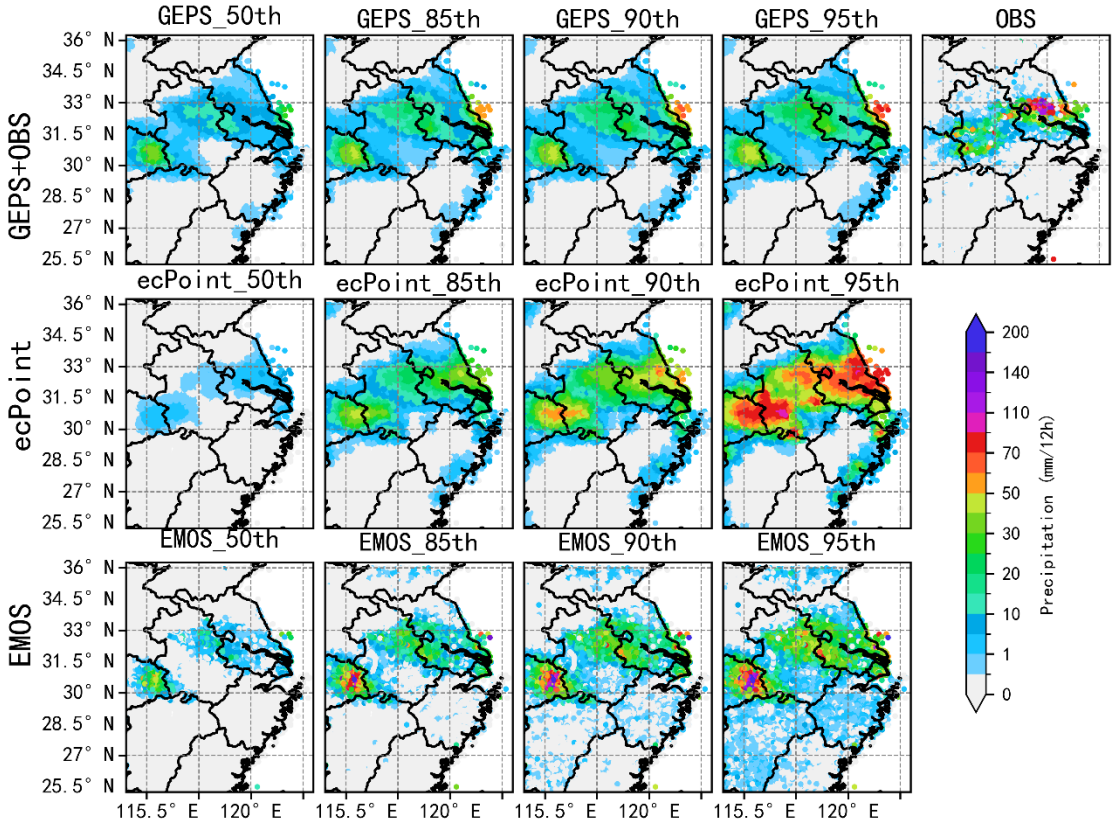

Fig. S4 Issued at 12:00 UTC on 6 July 2023, forecasting 12-hour precipitation for the 50th, 85th, 90th, and 95th percentiles valid for the 12-24h period. Observations are shown at 12:00 on July 7, 2023 (mm/12h).

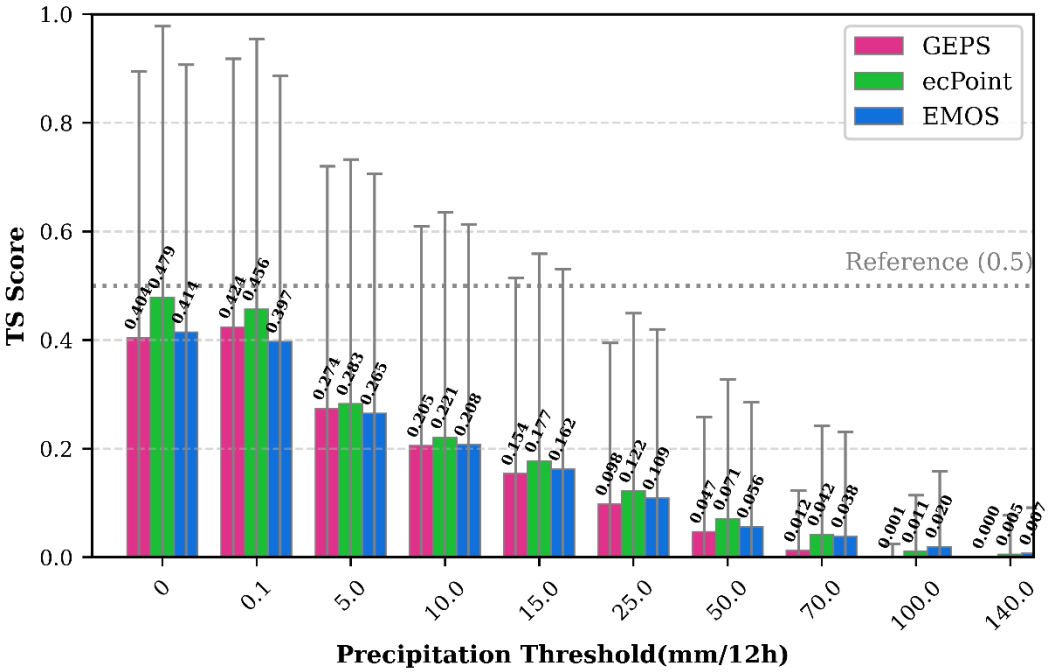

Fig. S5 TS Scores for the 75th percentiles of three ensembles under various thresholds, where annotated values indicate boxplot averages (mm/12h).

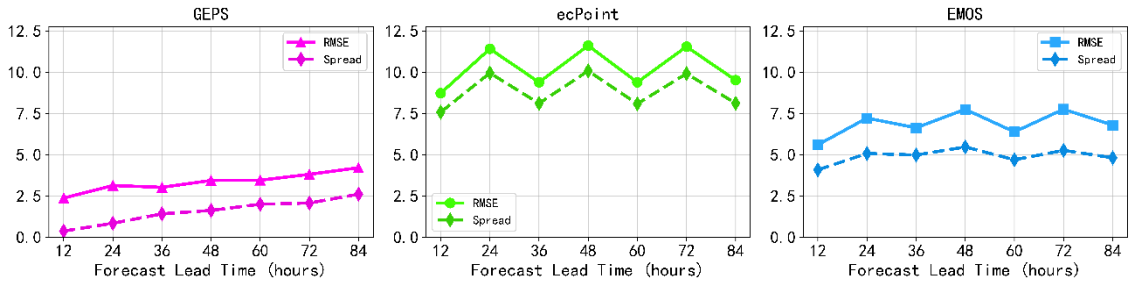

Fig. S6 TS Scores for the 75th percentiles of three ensembles under various thresholds, where RMSE and Spread of three ensemble means at different forecast leadtimes（mm/12h).