



Evaluating Short-Range Forecasts of a 12 km Global Ensemble Prediction System and a 4 km Convection-Permitting Regional Ensemble Prediction System

ASHU MAMGAIN,¹ S. KIRAN PRASAD,¹ ABHIJIT SARKAR,¹ GAURI SHANKER,¹ ANUMEHA DUBE,¹ and ASHIS K. MITRA¹

Abstract—Information regarding the uncertainty associated with weather forecasts, particularly when they are related to a localized area at convective scales, can certainly play a crucial role in enhancing decision-making. In this study, we discuss and evaluate a short-range forecast (0–75 h) from of a regional ensemble prediction system (NEPS-R) running operationally at the National Centre for Medium Range Weather Forecasting (NCMRWF). NEPS-R operates at a convective scale (~ 4 km) with 11 perturbed ensemble members and a control run. We assess the performance of the NEPS-R in comparison to its coarser-resolution global counterpart (NEPS-G), which is also operational. NEPS-R relies on initial and boundary conditions provided by NEPS-G. The NEPS-G produces valuable forecast products and is capable of predicting weather patterns and events at a spatial resolution of 12 km. The objective of this study is to investigate areas where NEPS-R forecasts could add value to the short-range forecasts of NEPS-G. Verification is conducted for the period from 1st August to 30th September 2019, covering the summer monsoon over a domain encompassing India and its neighboring regions, using the same ensemble size (11 members). In addition to standard verification metrics, fraction skill scores, and potential economic values are used as the evaluation measures for the ensemble prediction systems (EPSs). Near-surface variables such as precipitation and zonal wind at 850 hPa (U850) are considered in this study. The results suggest that, in some cases, such as extreme precipitation, there is a benefit in using regional EPS forecast. State-of-the-art probabilistic measures indicate that the regional EPS has reduced under-dispersion in the case of precipitation compared to the global EPS. The global EPS tends to provide higher skill scores for U850 forecasts, whereas the regional EPS outperforms the global EPS for heavy precipitation events (> 65 mm/day). There are instances when the regional EPS can provide a useful forecast for cases, including moderate rainfall, and can add more value to the global EPS forecast products. The investigation of diurnal variations in precipitation forecasts reveals that although both models struggle to predict the correct timing, the time phase and peaks in precipitation in the convection-permitting regional model are closer to the observations.

Keywords: Ensembles, forecast verification/skill, uncertainty, decision making, high impact weather, model evaluation/performance, regional models.

1. Introduction

The primary objective of forecasts is to mitigate uncertainty by providing valuable information to users. Weather details specific to location and time is critical for decision-making across various industries. Those include agriculture, hydrology, transportation, energy, construction, and defense services.

Parameterizing deep convection is a significant source of uncertainty in coarse-resolution models (Yano et al., 2018). Models that permit convection (CP models) prove advantageous for precipitation forecasting by explicitly resolving convection and capturing finer details of topography. Moreover, a CP ensemble prediction system (CP-EPS) has the capability to quantify uncertainty in forecasts over complex terrain. Over the past decade, advancements in supercomputer capabilities have empowered meteorological services to explore ensemble applications at CP scales (Clark et al., 2012; Golding et al., 2014; Klasa et al., 2018). The utilization of CP-EPS in many studies has identified added value particularly in predicting heavy precipitation events by dynamically downscaling global ensemble predictions (Weusthoff et al., 2010; Wang et al., 2011; Duc et al., 2013; Schellander-Gorgas et al., 2017; Gowan et al., 2018; Frogner et al., 2019; Schwartz, 2019; Wastl et al., 2021; Capecchi, 2021). However, even CP-EPS face challenges that require further improvement. Duc et al. (2013) identified shortcomings in predicting light rainfall in convective scales,

¹ Ministry of Earth Sciences, National Centre for Medium Range Weather Forecasting, Government of India A-50, Sector 62, 201 309 Noida, India. E-mail: ashu.mamgain@gov.in

indicating that regional models might not completely resolve convective cells. Whereas, Holloway et al. (2012) identified that parameterized convection models have tendency to generate light rain occurrences too frequently as compared to the observation. Gowan et al. (2018) demonstrated that the performance of CP-EPS could be affected by insufficient ensemble spread. Additionally, Frogner et al. (2019) noted challenges related to a decline in predictability for precipitation at scales smaller than ~ 60 km within the first 6 h in his experiments. Nevertheless, there is added value for both severe precipitation events and precipitation/no precipitation decisions for shorter lead times. Schwartz (2019) concluded from his experiments that the benefits of CP ensembles are primarily significant for forecast lengths up to 48 h.

These studies suggest that CP-EPS find applications in various contexts, and improvements in certain aspects, especially concerning small-scale forecasts compared to coarser model forecasts. However, it's important to note that the skill of convection-permitting models at longer lead times is known to be limited, and expecting a perfect point-to-point agreement between model forecasts and observations is unrealistic (Hohenegger and Schär, 2007). Weusthoff et al. (2010) further suggested that the traditional verification methods may prove inefficient due to their insistence on exact matches between forecasts and observations in terms of time and location, thereby overlooking the small-scale variability. Here, traditional ensemble verification methods comprise metrics like Brier Score, the Continuous Ranked Probability Score, the Rank Probability Score, the Reliability Diagram, and more. This may underscore the significance of enhanced value in CP models over finer grid-specific information. The added value of the forecast predictability also varies across seasons and locations. Also, the differences between the models in various studies are significant and robust against small changes in the verification settings.

Over the tropics, most rainfall originates from convective systems, making rainfall forecasting challenging, especially regarding intensity and diurnal timing in this region. There are a limited number of studies investigating the potential advantages of CP-ENS in the broader context of the tropics,

specifically for South Asia region. Maurer et al. (2017) investigated that a single model setup integrated with land surface and atmosphere perturbations, demonstrated higher skills in predicting precipitation than the multimodal setup over West Africa. They also highlighted the under dispersive nature of CP-ENS using a single model setup. For tropical East Africa region, Cafaro et al. (2021) found that CP-EPS is more skillful in predicting rainfall location and discriminating between events and nonevents. Comparing CP-EPS with a parameterized convection ensemble, Ferrett et al. (2021) revealed that the representation of convection plays a more significant role than grid resolution in experiments covering the Southeast Asia domain. Across different Indian regions, some studies in the past have utilized regional models with explicit convection. Ensembles were created with multi physics options (Kirthiga et al., 2021; Sisodiya et al., 2022), exploring the role of representing deep convection and orography in predicting precipitation events at the local scale. However, a comprehensive study to explore the scale dependence of forecast skill over the South Asian region is lacking, particularly in the context of added value with driving lower-resolution counterpart.

This study utilizes a unified framework to assess the role of explicitly resolving convection with finer topographical detail. The aim is to measure the effectiveness of CP-EPS forecasts for low-level wind and precipitation in the South Asian domain. Additionally, we further investigate the performance of CP-ENS in terms of its ability to capture the diurnal cycle of convective summer monsoon precipitation forecasts in the core monsoon zone. The global version of the EPS with a 12 km horizontal grid size, known as NEPS-G, has been operational at NCMRWF since June 1, 2018. Detailed descriptions of this high-resolution EPS implementation and its performance are discussed in Mamgain et al., (2018b, 2019, and 2020). Additionally, the regional Ensemble Prediction System with explicit convection (NEPS-R) has been operational at NCMRWF since July 2019. This marks the first time in India that an EPS is running operationally at a CP scale. Details about the model configuration are described in Prasad et al. (2019). This study focuses on the comparative analysis of the performances of NEPS-G and NEPS-

R at short range. While NEPS-G exhibits good skill in medium-range forecasting of large-scale atmospheric features, its coarse horizontal resolution and inability to resolve convective physical processes limit its skill in short-range forecasts at finer scales, particularly regarding the intensity of high-impact events. The goal of this study is to enhance users' intuitive understanding of the products from the NEPS-R, providing insights for future planning and improvements in the modeling system.

Different verification metrics have been used to assess the capabilities of NEPS-G and NEPS-R and the performances are evaluated extensively over two months, August and September of 2019 i.e., during the South Asian summer monsoon. The regional models are tuned with a focus more on specific weather phenomena close to the surface than at the higher vertical levels. Variables at lower atmospheric levels such as zonal wind at 850 hPa and precipitation over the Indian region are considered. This study will also give a measure of temporal variation of the forecast skill of NEPS-R from day 1 to day 3. The results thus can provide information to the forecasters about the capability and limitations of probabilistic forecast from a convection-permitting ensemble prediction system over the Indian domain.

The next section describes the characteristics of the EPSs and the observation data. Section 3 introduces the strategy and verification methods that we have used. The verification results and discussion based on the skill scores followed by the actual comparative verification are given in Sect. 4. Section 5 presents the comparison of diurnal variations in precipitation forecasts. Finally, summary of the results is provided in Sect. 6.

2. Model and Data

The global model NEPS-G provides the lateral boundary and initial conditions to the regional model NEPS-R. Both operational NEPS-G and NEPS-R are based on Met Office global and regional versions of the ensemble prediction systems known as MOGREPS. NEPS-G uses the configuration of the Unified Model based on Wood et al. (2014) and Walters et al. (2017). It comprises a total of 22

perturbed forecasts along with one control forecast. The 22 analysis perturbations, including horizontal wind speed components, potential temperature, specific humidity, and exner pressure, are generated through the ensemble transform Kalman filter (ETKF) method, utilizing forecast perturbations from previous cycles (Bishop et al., 2001). Perturbations are also applied to deep soil temperature, soil moisture content, and sea surface temperature, as outlined by Tennant and Beare (2014). Additionally, NEPS-G incorporates two stochastic physics schemes representing the effects of structural and subgrid-scale model uncertainties. Those physics schemes are random parameters (RP) and stochastic kinetic energy backscatter (SKEB) schemes (Bowler et al., 2008; Tennant et al., 2011). The perturbations in NEPS-G are added to the analysis fields prepared by the hybrid-4DVar method (Clayton, 2013; Kumar et al., 2018) to produce multiple perturbed initial conditions. Though 22 analysis perturbations are generated using ETKF at a 6-hourly cycle, only 11 perturbed initial conditions of 00 and 12 UTC are used for long forecast of 10 days forecast lead time. In NEPS-G, each perturbed member is considered to have an equal probability of occurring, meaning the system treats all generated scenarios with the same level of likelihood. This approach ensures an unbiased representation of potential outcomes. Operationally, the size of 11 perturbed members in NEPS-G is determined based on the optimal use of available resources. The eleven members, which run for 10 days forecast lead time from the initial condition of 00 UTC, provides the initial and boundary conditions to 11 ensemble members of NEPS-R.

NEPS-R has a horizontal grid resolution of nearly 4 km and it consists of 11 perturbed members plus one unperturbed control member. Operationally it runs once a day from the initial condition of 00 UTC and provides probabilistic forecasts up to 75 h on a domain centering over India (62° – 106° E; 6° S– 41° N) and this domain has been shown in Fig. 1. Model uncertainties in NEPS-R are represented by RP scheme. It is designed to account for the structural and subgrid-scale sources of model error. The perturbed and unperturbed initial conditions from NEPS-G are reconfigured for NEPS-R. The lateral boundary conditions to NEPS-R are provided at a 3-h

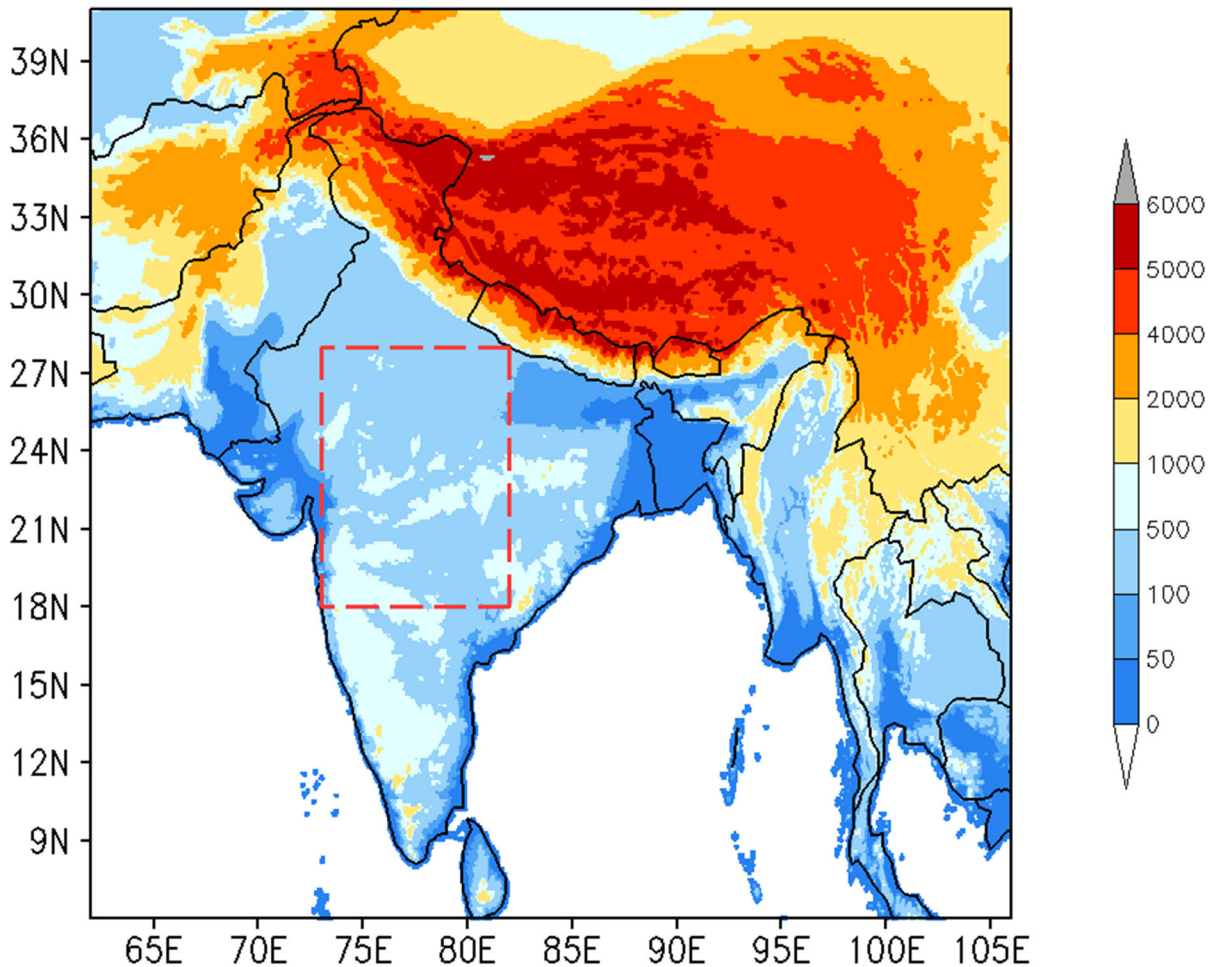


Figure 1

NEPS-R domain ($62\text{--}106^\circ\text{E}$; $6^\circ\text{S}\text{--}41^\circ\text{N}$) that has been used for the evaluation of statistics used in this study. Colour shading shows the model orographic height in the $\sim 4\text{ km}$ NEPS-R. The 'core monsoon domain' that is roughly considered over $73^\circ\text{--}82^\circ\text{E}$; $18\text{--}28^\circ\text{N}$ and marked with red dashed line box, was considered for the calculation of diurnal variation in precipitation and discussed in Sect. 5

frequency. NEPS-R has irregularly spaced but smoothly varying 80 hybrid-height levels with a lid at 38.5 km. These levels are terrain-following near the surface but relax towards the horizontal in the free atmosphere. Some more detail about the EPSs is summarized in Table 1.

For the verification of zonal wind at 850 hPa (U850), we have used the analysis data from the control member of NEPS-G. Here, the analysis is the initial condition or the best estimate of initial state that is used to initialize the deterministic forecast. In NCMRWF, the atmospheric data assimilation is originally running at a resolution of N320L70 that is

nearly at 40 km horizontal grid scale. We have used daily gridded rainfall data at $\sim 25\text{ km}$ resolution from merged products of satellite and gauge (Mitra et al., 2013). In this merged product, Global Precipitation Measurement (GPM; Hou et al., 2014) based satellite estimates are also used as a first guess for the rainfall. We have also used Integrated Multi-satellitE Retrievals for GPM (IMERG) precipitation from final run estimates available at half hourly intervals to investigate diurnal variations.

3. Verification Methods

The verification has been carried out over the domain of NEPS-R (Fig. 1) that is mainly covering the Indian and neighbouring region. Two important variables during monsoon months, zonal wind at 850 hPa (U850) and precipitation are considered for evaluation. We have used the standard verification metrics for the evaluation of NEPS-R with respect to NEPS-G. The metrics used here are the mean of ensemble-spread Vs ensemble root mean square error (RMSE) of the ensemble mean, rank histogram, area skill score (ASS) using the values of area under the relative operating characteristic (ROC) curve, brier score (BS), rank probability score (RPS), continuous rank probability score (CRPS), fractions skill score (FSS), reliability and sharpness diagrams and Potential economic value (PEV). Standard verification measures of the Global Ensemble Prediction System are based on the recommendation by WMO Manual on the Global Data-processing and Forecasting System, 2017 in its Appendix 2.2.35. Also, the diurnal variation in rainfall is compared using hourly data of observed rainfall. The methods are discussed in detail in the next section.

The forecast data of both models used for verification are projected to a common grid size. For rainfall evaluation, the forecast data were brought to the observation grid size that is ~ 25 km whereas for U850, the model grid of higher resolution that is NEPS-R was re-gridded to ~ 12 km, a coarser grid size of NEPS-G. The regional model at 4 km reso-

lution is expected to forecast rainfall at a higher intensity range. According to the criteria set by India Meteorological Department (IMD), rainfall amount in the range of 15–65 mm/day is considered to be in the “moderate” category. Values greater than 65 mm/day are used as a threshold limit for heavy rainfall cases. We have used both categories for rainfall verification. Values above the 95th percentile are also considered while calculating the fractions skill score of rainfall forecast. For U850, dichotomous events are selected based on one standard deviation greater than the sample data climatology.

4. Results and Discussion

In this section, we briefly use the popular approach of assessing probabilistic forecasts. Here, we also compare the growth in the forecast errors during first the 72 h of forecast predicted by both NEPS-R and NEPS-G for the same number of 11 ensemble members at 00UTC.

4.1. Spread Vs RMSE

The method of Spread-skill relationship is used here to check the extent of dispersion in NEPS-R and NEPS-G. This method has been widely used to evaluate the statistical reliability of EPSs (Johnson & Bowler, 2009). EPSs are generally under-dispersed as all the sources of uncertainty are not accounted for by the forecasting system. In a perfect case, when all the

Table 1

Main characteristics of the NEPS-R and NEPS-G

	NEPS-R	NEPS-G
Horizontal resolution (grid size)	4 km	12 km
Ensemble Size	11 members + 1control	11 members + 1control
Vertical levels	80 (Top:38.5 km)	70 (Top: 80 km)
Grid points	1200 × 1200	2048 × 1536
Domain	62° E-106° E; 6° S 41° N	Global
Forecast length	75 h	10.5 days
Model time step	2 min	5 min
Long forecast length	Once per day (00 UTC)	Twice per day (00 and 12 UTC)

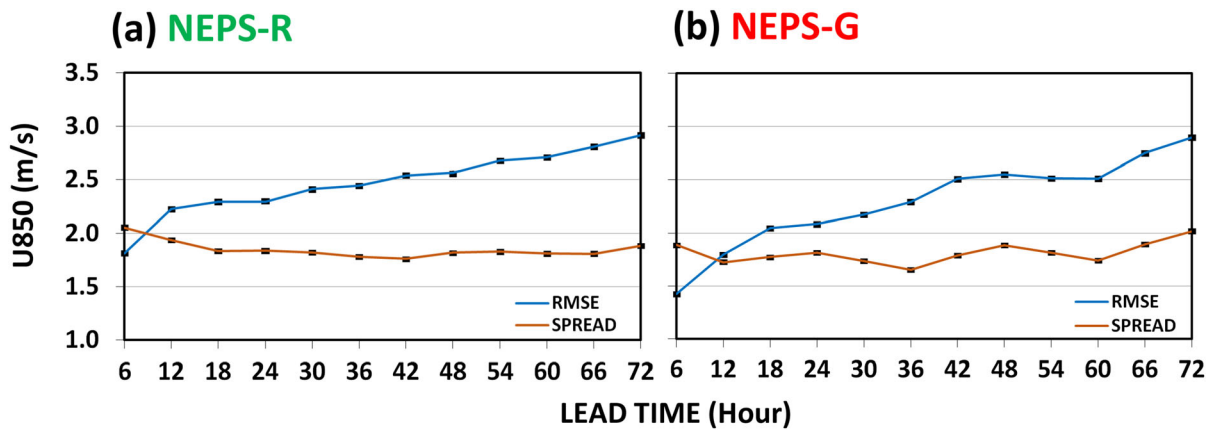


Figure 2

RMSE and ensemble spread of the **a** NEPS-R and **b** NEPS-G for zonal wind (m/s) at 850 hPa as a function of forecast lead time in hours. Error bars (black marks) indicate 95% confidence interval using the bootstrap method

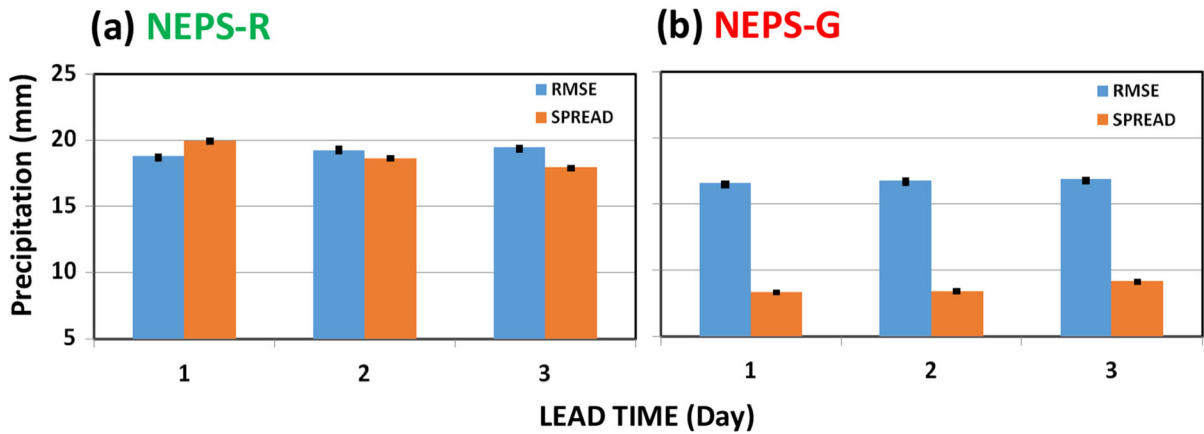


Figure 3

Bar diagram of RMSE and ensemble spread of 24 h accumulated precipitation (mm/day) for **a** NEPS-R and **b** NEPS-G as a function of forecast lead time in days. Error bars (black marks) indicate 95% confidence interval using the bootstrap method

sources of uncertainties associated with the analysis data and model physics are represented by the EPS, then the RMSE of ensemble mean and ensemble spread curves can match each other (Palmer et al., 2006).

For the NEPS-R domain, averaged RMSE of the ensemble mean and average spread of the ensembles are plotted as a function of forecast lead time and shown in Fig. 2 for U850 and Fig. 3 for precipitation. A re-sampling technique known as bootstrap method has been used to estimate the statistical significance of spread and RMSE scores at 95% confidence interval as shown with error bars. For U850, the

forecast and analysis data both are available at 6 hourly intervals. In the case of precipitation, the forecast of 24 h of accumulated rainfall is verified with quality-controlled daily satellite-gauge merged precipitation data. Here, the average spread is calculated as the square root of the averages over all forecasts of the ensemble variance. In the case of NEPS-G, growth in the ensemble spread includes the model error through the stochastic physics scheme and the inflation factor. An inflation factor acts as a tuning variable that adjusts the ensemble spread to the ensemble mean error. It inflates the analysis perturbation amplitude to adjust the

ensemble forecast variance consistent with the unperturbed forecast error variance. We can see in Fig. 2b that the ensemble spread of NEPS-G is varying between 1.6 and 2 (m/s) and not increasing with forecast lead time. This may be due to the small size of the ensemble that is 11 in this study. Similarly, in NEPS-R (Fig. 2a), the spread in U850 is slightly higher during starting hours only and after that, it nearly remains constant as it is seen in the case of NEPS-G. Here the small error bars indicate that the results are not overly sensitive to small changes in the data and hence significant. RMSE is slightly better in the case of NEPS-G and that can be attributed to the error computation with respect to global analysis with coarser resolution. Since the values of error bars are not overlapping, significant difference is noticed in the case of 24 h accumulated precipitation. Figure 3 shows that both RMSE and spread are much closer to each other in NEPS-R (Fig. 3a) as compared to NEPS-G (Fig. 3b). A larger spread is noticed in NEPS-R and the precipitation values are in between 18 and 20 (mm/day). In the case of NEPS-G the precipitation values are comparatively more under-dispersed. There is a slight decrease in spread per day in the case of NEPS-R, whereas NEPS-G shows a slight increase in precipitation spread with forecast lead time. RMSE is slightly better in the case of NEPS-G. In this case also, the observation data is at a coarser grid (~ 25 km). Although precipitation observation is quality controlled, it is also possible that the coarse resolution observations have smoothed precipitation and it affects particularly higher intensity range. In a previous study by Mamgain et al. (2018a), a number of rainfall days with very heavy and extremely heavy categories are overestimated in the control version of NEPS-R in comparison to the observation. The uncertainty in verifying analysis is well known and has effect on the verification statistics (Bowler, 2008; Candille & Talagrand, 2008). Further, for the shorter lead time, the forecast has less time to diverge from the actual conditions and therefore any discrepancy can primarily arise from the uncertainties and limitations in the observed data. Therefore it is important to consider uncertainty in the observation while interpreting the verification scores.

4.2. Rank Histogram

The rank histogram is also known as the Talagrand diagram or binned probability ensemble (Anderson, 1996; Hamill, 2001). It represents the rank frequency of the verifying observation/analysis relative to the values of ensemble members sorted from lowest to highest. The observed probability distribution is expected to be well represented by the members of the ensemble. By observing the shape of the histogram, the nature of bias and spread in the ensemble system can be understood. The uniform rank distribution of the EPS indicates a reliable system but a flat distribution may also be generated from unreliable ensembles. Here the uniformity of a rank histogram has been tested using Pearson chi-square (χ^2) goodness-of-fit test under the null hypothesis of a flat rank histogram (Jolliffe & Primo, 2008; Wilks, 2019). For the large sample size and under-dispersed ensembles, the χ^2 test is considered to be a powerful approach.

An under-dispersive ensemble system has a U shape rank histogram whereas a bell-shaped distribution indicates an over-dispersive system. The shape of the rank histogram in Fig. 4 for U850 indicates that both NEPS-R (Fig. 4a–c) and NEPS-G (Fig. 4d–f) are under-dispersive or over-confident forecasting systems in all forecast lead days. In the case of NEPS-G, larger populations at the lower ranks mean forecast have a positive bias or over-forecasting. NEPS-R shows that the ensemble has little spread. Figure 5 shows the rank histograms for day 1, day 2, and day 3 precipitation forecasts by NEPS-R (Fig. 5a–c) and NEPS-G (Fig. 5d–f). While NEPS-R shows dry bias which increases with forecast lead time NEPS-G exhibits tendency of overestimating precipitation. The positive precipitation bias in the global deterministic version due to overestimated frequency of light precipitation events is discussed in Mamgain et al., (2018a). It is also discussed that the convection-permitting model has the tendency to underestimate statistics of light rainfall events. Since the sample size is very large, the goodness-of-fit test for the flatness of rank histogram is highly significant. Although the χ^2 statistic are large as shown in Figs. 4 and 5, U850 χ^2 statistic are smaller in NEPS-G as compared to NEPS-R (Fig. 4) and those are

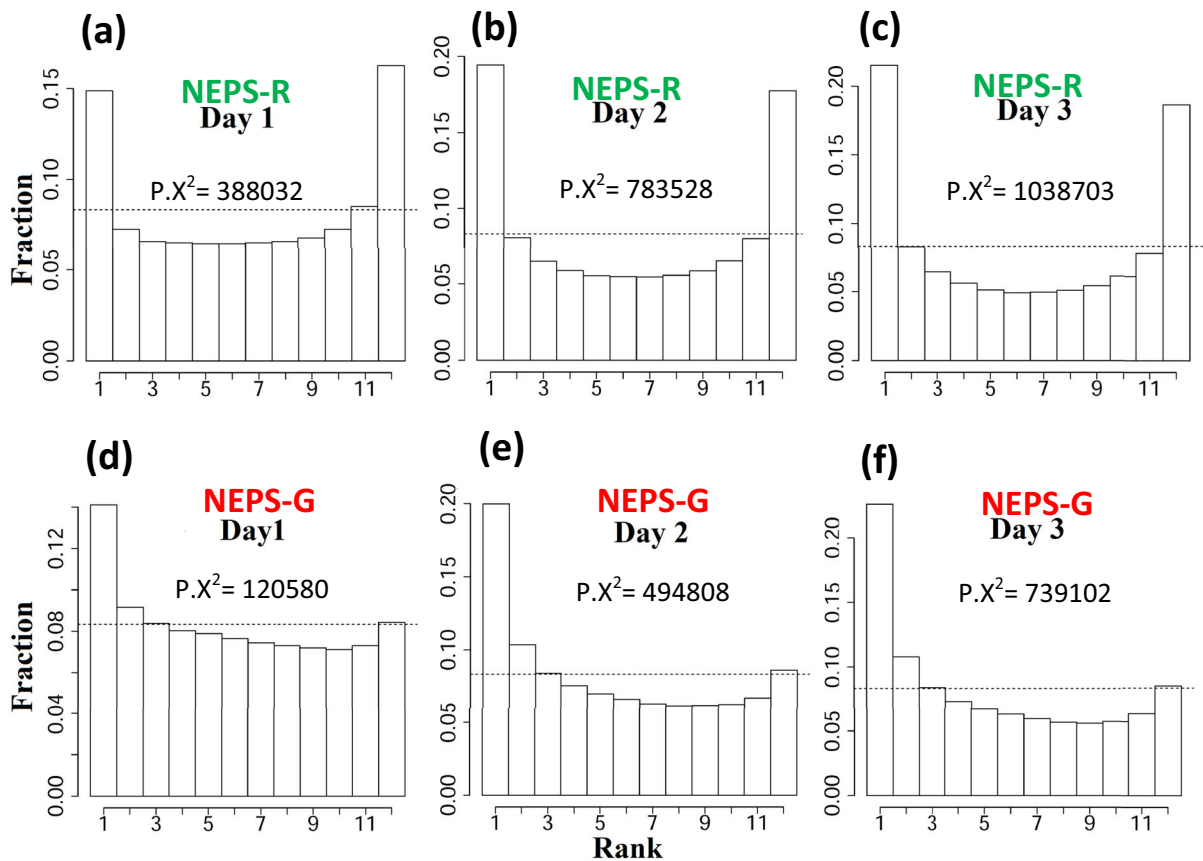


Figure 4

Rank histograms for Day 1 (a, d), Day 2 (b, e) and Day 3 (c, f) forecasts of NEPS-R (a, b, c) and NEPS-G (d, e, f) for zonal wind at 850 hPa. Pearson Chi-Square test statistic that could indicate deviation from flatness of a rank histogram is shown here

increasing with forecast lead time. Whereas for precipitation in Fig. 5, NEPS-R shows lower χ^2 statistic means more uniformity in rank distribution as compared to the NEPS-G.

Verification scores such as area skill score, brier score, continuous rank probability score, and rank probability score are shown in Fig. 6 for U850 and Fig. 7 for precipitation. Following is the point-wise discussion based on matrices used for measuring the skill of the forecasts.

4.3. Area Skill Score (ASS)

The relative (or receiver) operating characteristic (ROC; Mason, 1982) curve represents a variation of hit rate with false alarm rate for threshold probability ranges between 0 and 1. It is the measurement of the forecasting system's ability in discriminating

between events and non-events. The higher the area under the ROC curves (AUC) better is the capability to distinguish between event and non-event cases. An AUC value equal to 0.5 indicates that the model has no skill. Scores in the middle of 0 to 1 are hard to interpret as good or bad and so sometimes skill scores are calculated. The ASS can be calculated using the values of the AUC (Richardson, 2000). The improvement in the forecast can be determined by comparing it with the unskilled prediction where the hit rate is equal to the false alarm rate. The value of AUC in the case of the unskilled forecast is 0.5 ("random chance"). It is considered here as a reference forecast. So the result indicates how good the forecast is in terms of % improvement in score compared to the reference forecast with no skill. The ASS can be defined as:

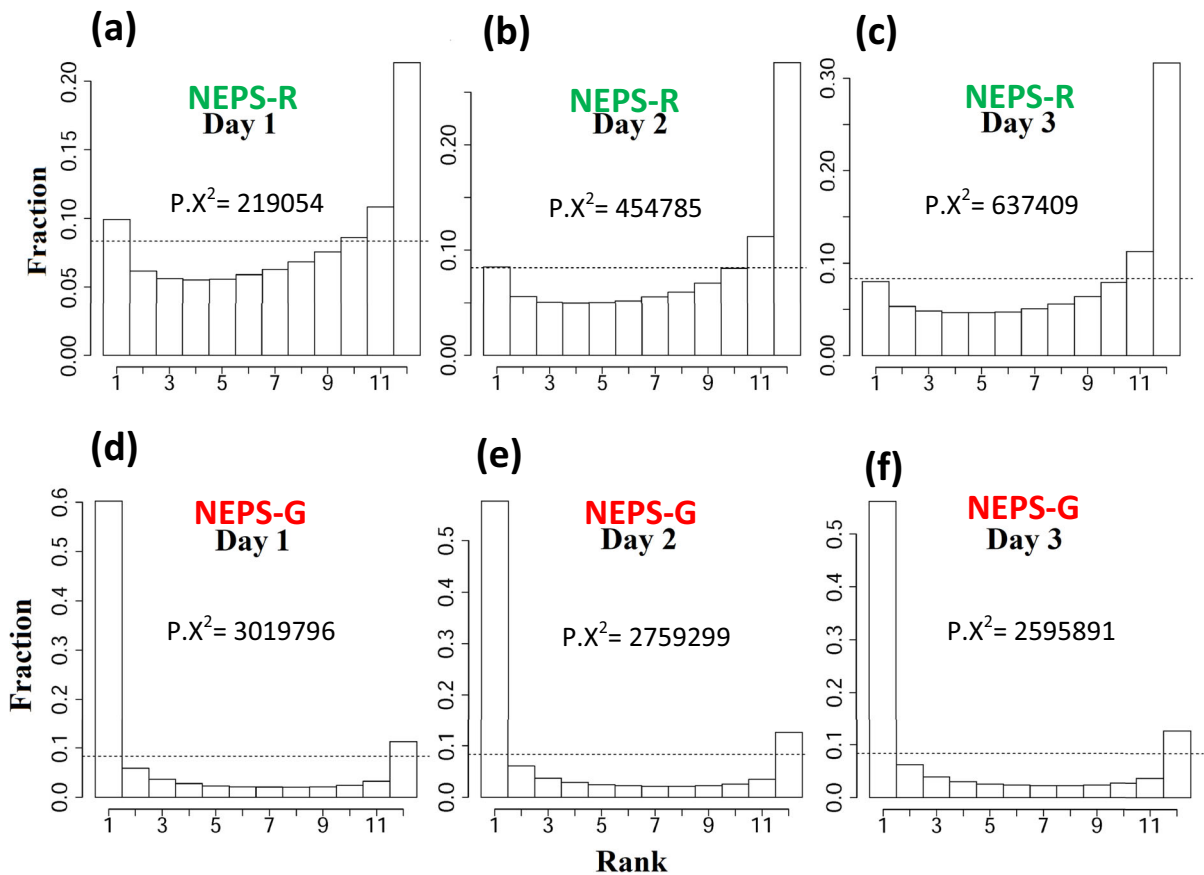


Figure 5

Rank histograms for Day 1 (a, d), Day 2 (b, e) and Day 3(c, f) forecasts of NEPS-R (a, b, c) and NEPS-G (d, e, f) for 24 h accumulated precipitation. Pearson Chi-Square test statistic that could indicate deviation from flatness of a rank histogram is shown here

$$\begin{aligned}
 ASS &= \frac{AUC_{forecast} - AUC_{unskilled}}{AUC_{perfect} - AUC_{unskilled}} \\
 &= \frac{AUC_{forecast} - 0.5}{1 - 0.5} \\
 &= 2AUC_{forecast} - 1
 \end{aligned}$$

The significant testing for the difference between AUC values of both the EPSs has been performed using bootstrap method and found that the true difference in AUC is not equal to zero. The ASSs of the models are calculated using AUCs and shown in Fig. 6 afor U850 and Fig. 7a and b for precipitation. A positive value of a skill score indicates an improvement or skillfulness in the forecast relative to the reference forecast or sample climate. U850 in Fig. 6a shows that the discrimination property of both

models has skill above 0.6 till 72 h of the forecast. The NEPS-G in the case of U850 has shown a better skill score as compared to the NEPS-R. As discussed earlier also, the assessment of forecast skill for U850 has been conducted using the respective models' analyses. In the NEPS-G model, the preparation of the verifying analysis through data assimilation (DA) techniques is expected to provide the best possible estimation of the atmospheric state. DA methods integrate background information from the model with available quality-controlled observations. However, analysis data derived from DA methods at a global scale may have limitations in capturing fine-scale details, especially in regions characterized by complex topography and sparse observational coverage. Additionally, as outlined in Sect. 2 of this study, DA in NEPS-G operates at approximately a 40 km

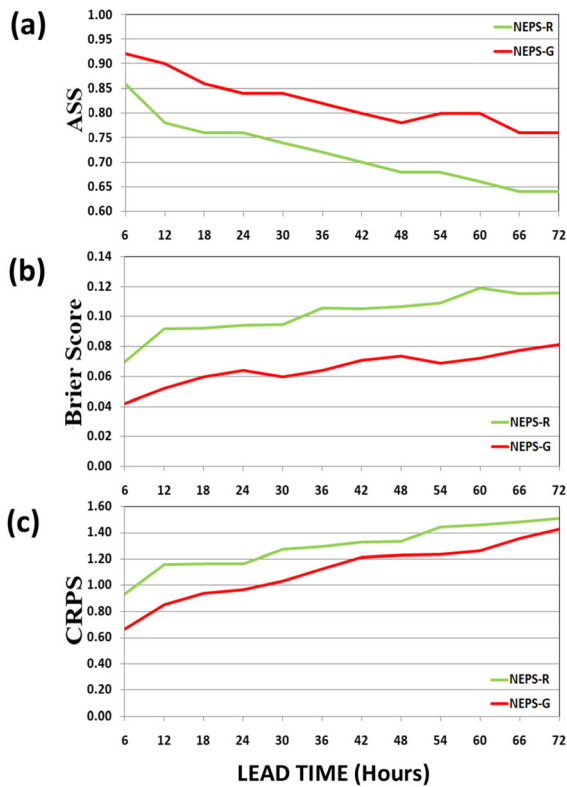


Figure 6

Verification skill scores **a** ASS, **b** brier score and **c** CRPS of NEPS-R and NEPS-G for zonal wind (m/s) at 850 hPa as a function of forecast lead time in hours. For ASS and Brier Score calculations, zonal wind greater than 1 standard deviation from the sample climatology is considered

horizontal grid scale. In contrast, convective-scale data assimilation in NEPS-R is still under research and is beyond the scope of the present study. High resolution NEPS-R with explicit convection can plausibly show the ability to capture fine-scale features due to the improved representation of orography. However, for validation of the results, observational data at a similar grid scale is a key factor.

In the case of precipitation, greater than 15 mm per day in Fig. 7a, ASS is above the value of 0.4 till day 3 forecasts in both models. It is interesting to see that the skill of NEPS-R did not deteriorate for higher rainfall intensity (> 65 mm per day) in Fig. 7b. However, NEPS-G which performs better for light and moderate rainfall intensity events is not better than NEPS-R for heavy rainfall cases.

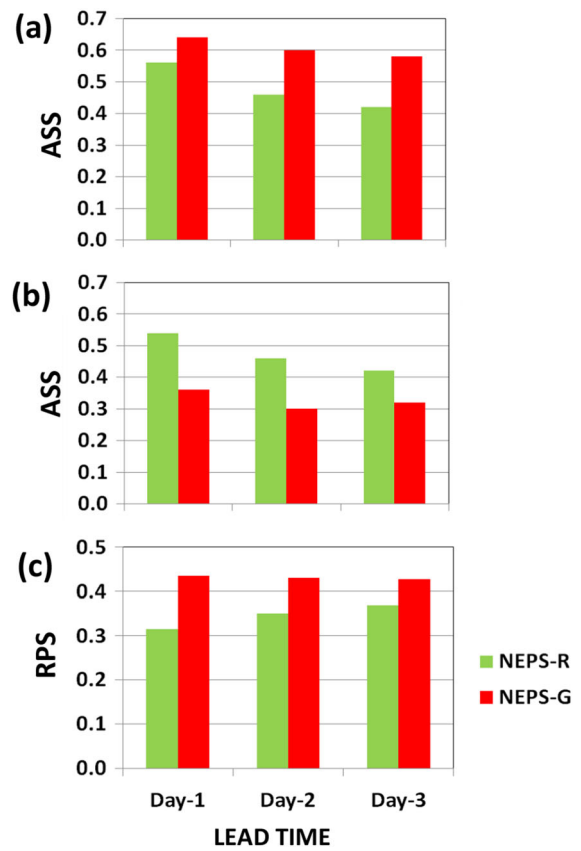


Figure 7

Verification skill scores of NEPS-R and NEPS-G are **a** ASS for the precipitation values greater than 15 mm/day, **b** ASS for the precipitation values greater than 65 mm/day and **c** RPS for the different rainfall categories (as explained in the text) as a function of forecast lead time in days

4.4. Brier Score (BS)

A popular verification score that concisely summarizes the concepts of reliability and resolution is the Brier Score (Brier, 1950). The Brier Score is a measure of probability forecast accuracy for dichotomous events. It is the average square deviation between the forecast probabilities and their outcomes and is defined as:

$$BrierScore = \frac{1}{N} \sum_{i=1}^N (p_i - O_i)^2$$

where p_i is the forecast probability which is the fraction of members predicting the event and O_i is the observed outcome with the value of 1 if an event occurs and 0 if an event does not occur. N is the

forecast–observations pairs considered. BS is negatively oriented. The best possible value of the Brier Score for total accuracy is 0 and the worst value is 1. It means scores closer to 0 indicating better forecast. In Fig. 6b, Brier scores indicate that the NEPS-G is performing better than NEPS-R for the U850 forecast for all 3 days' lead times. As previously mentioned, this could be due to the coarse resolution of the analysis data.

4.5. Continuous Rank Probability Score (CRPS)

The CRPS is analogous to Brier Score with an infinite number of continuous classes (Hersbach, 2000). It can be interpreted as the integral of the Brier score over a continuum set of all possible thresholds of the selected variables. The CRPS reduces to the mean absolute error in the case of a deterministic forecast. Like the Brier score, CRPS is also negatively oriented. For a perfect deterministic system, it reaches a minimum value of zero. A lower value of the CRPS indicates that EPS has better skills. Figure 6c indicates that NEPS-G is more skillful than NEPS-R for U850 at all forecast ranges.

4.6. Rank Probability Score (RPS)

The RPS is also an extension of the Brier Score but for a sum over a set of selected discrete forecast categories. It is defined as:

$$RPS = \sum_{k=1}^K \left[\left[\sum_{i=1}^k p_i \right] - \left[\sum_{i=1}^k O_i \right] \right]^2$$

where K is the number of forecast categories. $\left[\sum_{i=1}^k p_i \right]$ is the cumulative probability assigned by the model to the k^{th} component and p_i is the forecast probability in category i , $\left[\sum_{i=1}^k O_i \right]$ is the cumulative observation with O_i equals 1 if the true outcome (observation) falls in category i , and equals 0 otherwise. (Wilks, 2005). As discussed earlier in the text, the ASS is a metric used to evaluate the discrimination ability of a model in binary classification of events. In simpler terms, the ASS provides insight into the model's resolution, or its capability to accurately depict the differences between the categories such as "event" and "non-event". On the other

hand, RPS assesses the accuracy of probabilistic forecasts by comparing the ranked probabilities across different categories (it can be considered as a measure of bias as it is the difference between forecasts and observations). Unlike the ASS, which focuses on specific events or thresholds, RPS focuses on the entire probability distribution and quantifies the spread between the forecasted and observed probabilities for different categories or bins.

In the case of RPS as well, a smaller score means better skill. This method rewards the sharp forecast and emphasizes accuracy by penalizing large errors compared to near missed forecasts. Both models are verified against the same observation interpolated to a grid size of 25 km. We have used 0, 2, 15, 65, 115, and 195 mm per day rainfall categories for the calculation of RPS. The results based on RPS in Fig. 7c suggest that the NEPS-R can perform well in terms of probability distribution calibration as indicated by a lower value of RPS compared to NEPS-G. However, it possesses lower discriminating capabilities particularly for moderate rainfall amounts (rainfall > 15 mm/day) as indicated by a lower value of ASS (Fig. 7a). On the other hand, for heavier precipitation (rainfall > 65 mm/day), NEPS-R demonstrates better discrimination between event and non-event classification compared to NEPS-G (Fig. 7b).

4.7. Fractions Skill Score (FSS)

In the convection-permitting model, we are more interested in small-scale weather details. However, at this scale, local details tend to be noisier and contribute significantly to shorter lead time errors. Forecast errors associated with the convection-permitting model tend to grow more rapidly compared to the coarse resolution model. At the convective grid scale, model skill is often affected by the classical 'double penalty effect' or displacement error. Displacement error is a type of representation error that occurs when the model misplaces or misaligns atmospheric features compared to their observed positions. At the higher grid resolution, the main improvement can be seen in the reduction of representativeness error. That can only be verified against the observations for a specific location. The

availability of high-resolution quality-controlled observations at the surface level is very limited. FSS (Roberts & Lean, 2008) is a neighborhood verification method and does not require matching the fine-scale forecast exactly to the location of the observation grid. It is generally used for the verification of precipitation forecasts from numerical weather prediction models to determine the forecast accuracy as a function of spatial scales (Mittermaier et al., 2013; Weusthoff et al., 2010). It enables the comparison of forecasts of different grid resolutions with respect to the same observation. Also, we can define a minimum spatial scale at which a forecast can be considered useful or skillful. In this method precipitation exceeding a threshold value is used to compare fractions of model forecast and observed value within a selected domain. The value of FSS increases with an increase in fractions rainfall coverage. In most cases, an FSS value greater than 0.5 is considered a good sign of a useful forecast. In this method, the ensemble mean is used to make dichotomous events and the spatial distribution of events in a selected window is calculated probabilistically (Mittermaier & Csima, 2017).

Data from both models and observation are first converted to binary fields based on thresholds 15 and 65 (mm/day) accumulation and 95th percentile based on sample climatology. Here we are not considering the light rainfall intensity. In a previous study by Mamgain et al. (2018a) using a deterministic version of a regional model, it was noticed that at convection-allowing scale, the model underestimates the light rainfall events which are considered as a general tendency of convection-allowing models. If the data at grid points are found greater than the selected thresholds, they are given a value of 1 else the value is 0. The accumulation threshold value can be used for a one-to-one measure of model skill against observation whereas the percentile threshold will remove the impact of systematic biases in the rainfall forecast due to different grid resolutions.

In a selected window, if the forecast has the same frequency of events as those in the observation, then FSS is considered perfect and it scores 1. The FSS is computed as introduced by Roberts and Lean (2008):

$$FSS = 1 - \frac{MSE}{MSE_{ref}}$$

Here, the mean square error (MSE) and its reference (MSE_{ref}) are given by:

$$MSE = \frac{1}{N} \sum_{i=1}^N (O_i - F_i)^2$$

$$MSE_{ref} = \frac{1}{N} \left[\sum_{i=1}^N O_i^2 + \sum_{i=1}^N F_i^2 \right]$$

where the O_i and F_i are the observed and forecast fractions respectively at each point and N is the total number of grid boxes of the sliding window. MSE_{ref} is the lowest skilled reference forecast obtained from observed and forecast fractions. FSS can also be explained as the variations on the brier skill score and computed as fractions brier score divided by the mean of the sums of squared observed and forecast fractions.

The FSS is computed for all the required neighborhood sizes and the window size moves up to the whole domain. The variations in forecast skill as a function of neighborhood size are shown in Fig. 8. Precipitation accumulation values 15 mm/day and 65 mm/day are used as a threshold in Figs. 8a and b, respectively whereas Fig. 8c is based on percentile threshold (95th percentile). Other than the FSS variations a difference between the results obtained from these thresholds can be noted at the extreme right corner of these figures where the neighborhood size is covering the whole domain and FSS is reaching to a score 1. Accumulation based threshold shows some bias as FSS not indicating a perfect score of 1 over the whole domain. This difference is larger for a higher accumulation threshold which is 65 mm/day. Figure 8c indicates that percentile-based FSS has a score of 1 for the whole domain. The smallest scale at which the forecast is sufficiently skillful is (indicated by the dashed line) defined as $FSS > 0.5 + b/2$, where b is the observed fractions rainfall coverage of the sliding window (Roberts & Lean, 2008). In this case for (Fig. 8a) $b = 13\%$, (Fig. 8b) $b = 2\%$ and (Fig. 8c) $b = 5\%$. The value of 'b' is large for Fig. 8a because a large fraction of the model domain could be covered by moderate rainfall. The fraction values just above the dashed lines in

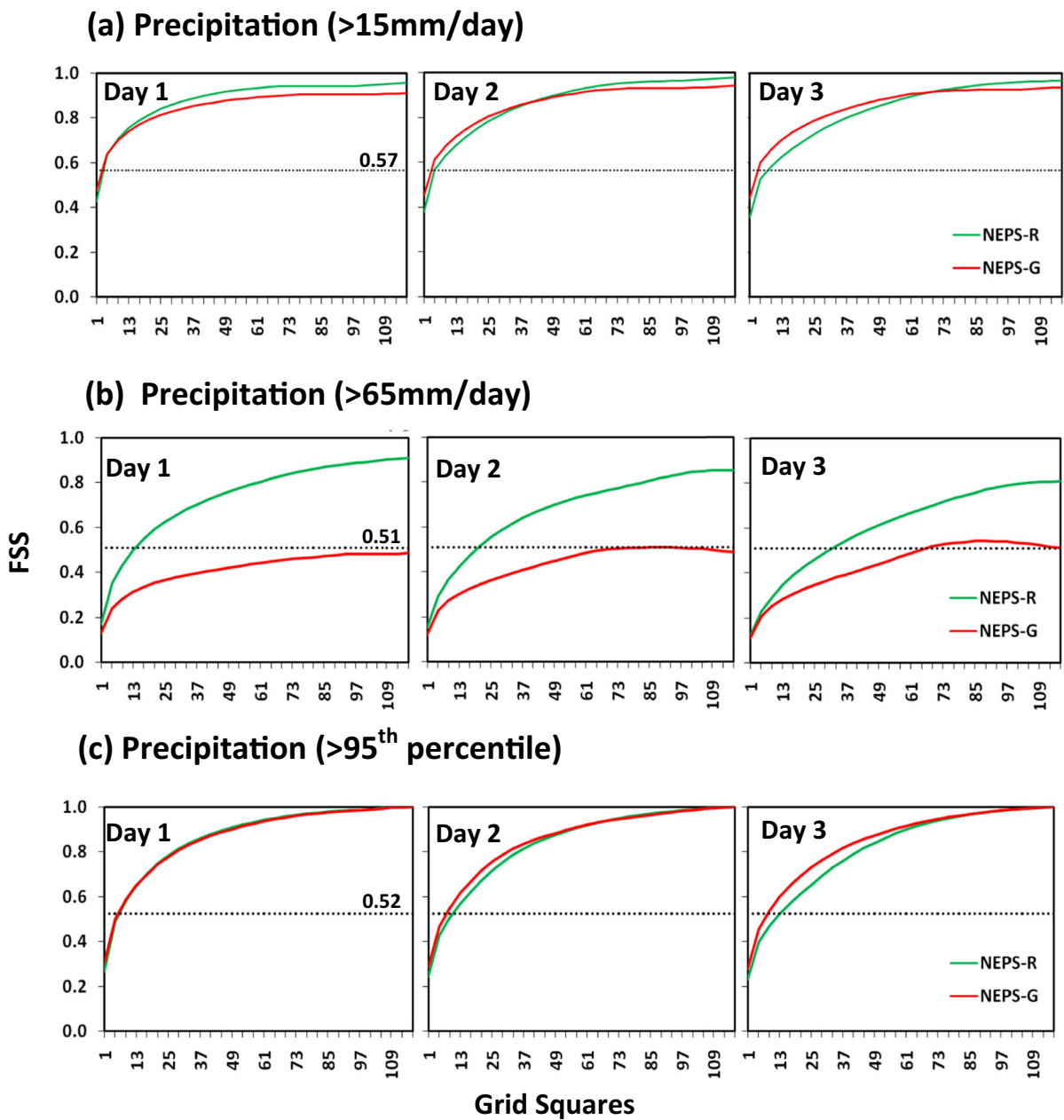


Figure 8

Fractions Skill Scores for day 1, day 2 and day 3 forecasts of 24 h accumulated precipitation greater than **a** 15 mm/day, **b** 65 mm/day and **c** 95th percentile of the sample climatology. The horizontal dashed black line indicates skill greater than a random forecast. Each grid size is equivalent to 25 km which is the resolution of observed precipitation

Fig. 8 are the values of FSS above which the forecast is skillful. Since the resolution of observed data is 25 km, the grid square of FSS represents the same resolution as both models first bring to this grid size. FSS for precipitation greater than 15 mm/day in

Fig. 8a depicts that day 1 forecast of NEPS-R has better scores over NEPS-G. On days 2 and 3, for smaller grid distance, NEPS-G has a better score but NEPS-R has higher scores for increased grid distance. Both models' forecast skill is above the reference line

roughly just after the 5 grids on day 1 and 13 grids on day 3. When the precipitation threshold is increased to 65 mm per day (Fig. 8b), NEPS-R outperforms its global counterpart for all the days' forecast. However, both models exhibit poor scores for small grid distance due to the presence of large spatial errors. Also, the fractions coverage bias of the precipitation is lower in the case of NEPS-R. NEPS-G does not improve much with increased grid distance and remained below the referenced line which indicates substantial high precipitation bias. The percentile-based threshold that is shown in Fig. 8c is based on the model climatology of sampled data. As the range of the simulated model forecast may not match the range of the observations, an approach based on percentile normalizes the scores for the simulated values that may not be as high as the observations. In both models we can notice (Fig. 8c) much-improved scores with removed bias. This enables a fairer assessment of scores for similar thresholds. The normalized fractions scores indicate better skill in day 2 and day 3 forecasts of NEPS-G for smaller grid size. For the day 1 forecast, the FSS curves of both models are overlapping.

The NEPS-G demonstrates the ability to better capture the signature of extremes when evaluated with respect to the sample climatology. On the other hand, NEPS-R may offer better skill compared to NEPS-G for extreme precipitation when the forecast is based on a threshold approach. The percentile-based method indicates the higher rate of reduction in the predictability in NEPS-R with lead time as compared to its global counterpart. Additionally, for moderate rainfall cases NEPS-R has the tendency of providing higher skill till day 1. However, for higher lead time the predictability of NEPS-R reduces for moderate cases compared to NEPS-G. Presently, high-resolution climatology is not available for the NEPS-G which could provide quality products based on the percentile-based approach. Re-forecasting the model climate for ensembles at high resolution is challenging due to its requirement of computation resources.

4.8. Reliability and Sharpness Diagrams

A reliability diagram can show the reliability and resolution components of an EPS by plotting the observed relative frequency as a function of its forecast probability (Wilks, 2005). It is a measure of agreement between the probabilistic forecasts and the observations. The diagonal line in this plot means the observed frequency matches with the forecast probability and the system is perfectly reliable. In the cases when the reliability curve falls below the diagonal line, the system over-forecasts and if the curve lies above then it under-forecasts. The sharpness diagram indicates how often the probabilities were issued and related sampling issues.

Figure 9 represents the reliability diagram for U850 and Fig. 10 is for the precipitation forecasts. Both Figs. 9 and 10 show that with an increase in the forecast probability of the occurrence of events, the verified chance of observing the event is also increasing. The reliability curves depict mostly over-forecasting. However, in both models, slight under-forecasting is also noticed in the lower-quintile category of U850 (day 2 and day 3 forecasts) as shown in Fig. 9b–e, and f and precipitation greater than 15 mm/day (Fig. 10). U850 reliability diagram (Fig. 9d–f) of NEPS-G shows more reliable curves compared to NEPS-R (Fig. 9a–c). Precipitation greater than 15 mm/day as simulated in NEPS-R (Fig. 10 a, b, and c) shows more reliability than NEPS-G on all the forecast days. However, NEPS-G is slightly better than NEPS-R for low precipitation forecast probability (< 20%). For the precipitation category greater than 65 mm/day, NEPS-R is only competing with NEPS-G at highest forecast probability which is close to 90%. Below that probability, NEPS-G shows better reliability. For high-resolution forecast, we can have better sharpness but reduced reliability and low probability due to the double penalty effect. Sharpness diagrams in both models (Figs. 9 and 10) show that most of the forecasts predict a lower probability than the climatological probability. Although, both models can predict probabilities slightly greater than 40% of the cases of precipitation forecasts (Fig. 10), however, the sample size used in this calculation are comparatively much

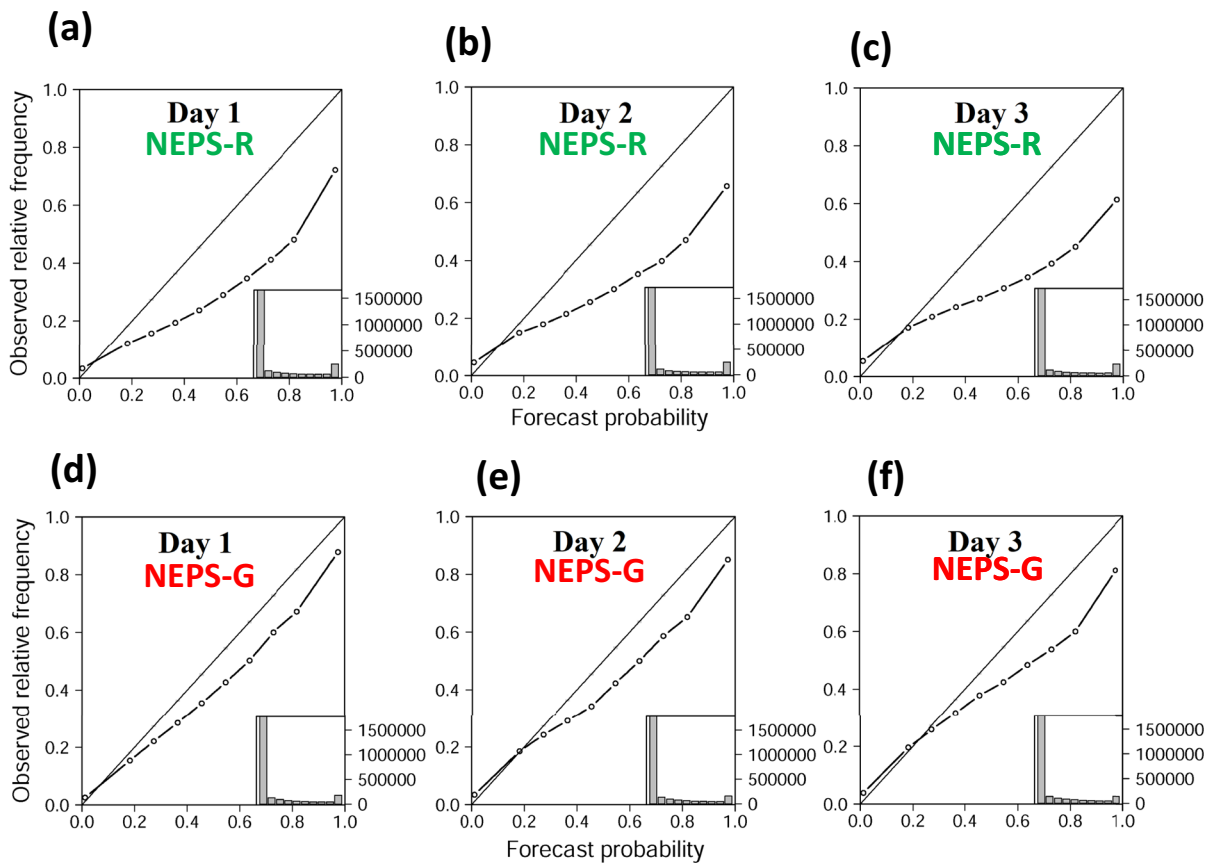


Figure 9

Reliability and Sharpness diagrams for day 1 (a, d), day 2 (b, e) and day 3 (c, f) forecasts from NEPS-R (a, b, c) and NEPS-G (d, e, f) of zonal wind at 850 hPa. Zonal wind greater than 1 standard deviation from the sample climatology is considered as event. The sharpness histogram represents the relative frequency of events in each forecast probability bin

lower than the sample size used in calculating models' skill for lower quartile category.

Using the most standard verification measure we have assessed different attributes which generally contribute to the quality of the forecasts. These aspects of forecast performance can strongly influence the value of a forecast. The forecast value can guide a decision maker to understand a level of benefit or loss while using the forecast products. Next, we will discuss this measure of forecast value that could contribute to decision making.

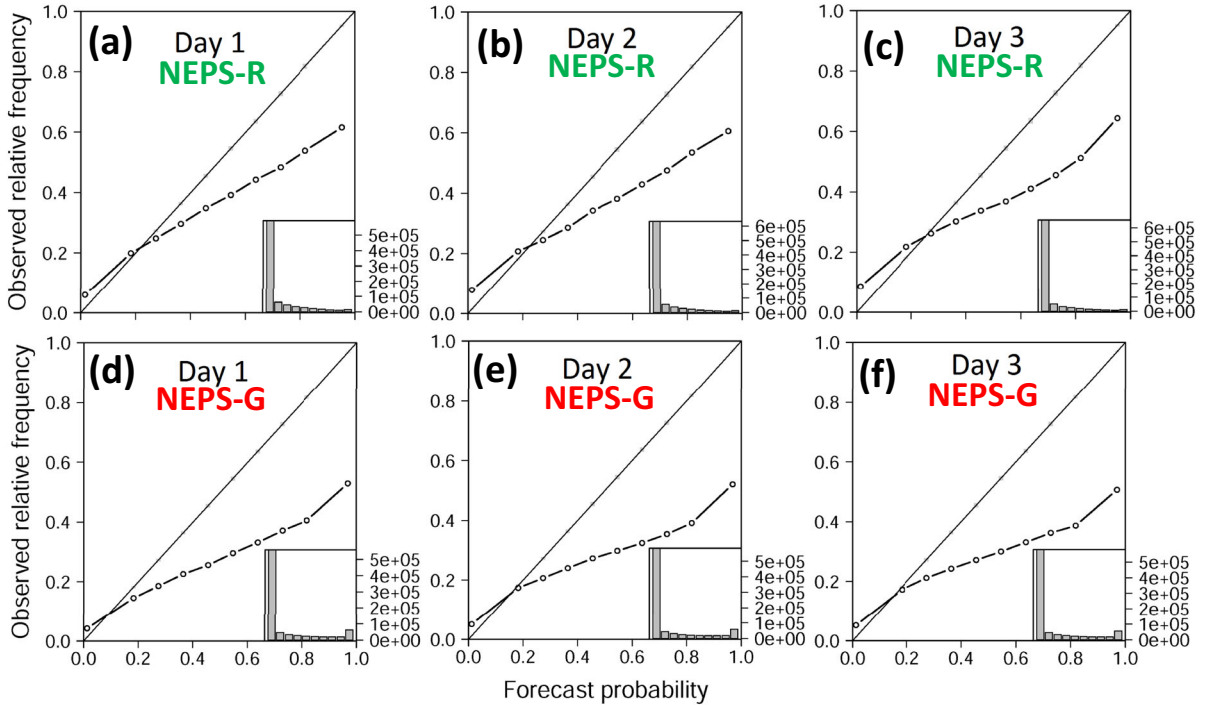
4.9. Relative or Potential Economic Value (PEV)

There are chances that forecasts are not of good quality but it might help the forecasters in providing

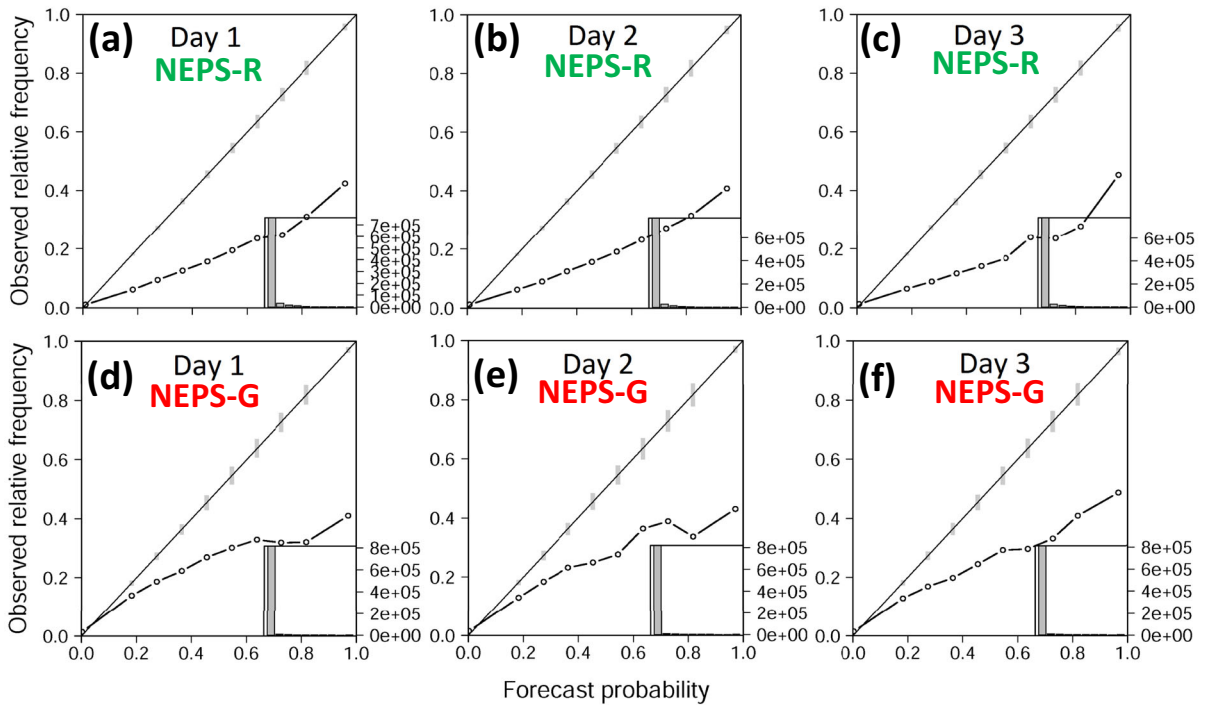
some useful information which is necessary to provide mainly when the cost of the missed event is high. The PEV is a skill score of the expected expense relative to the climatological details. The value of a system can be calculated using a cost-loss ratio (C/L) decision model (Richardson, 2000). Here C is the cost of taking preventive action and L is the loss if no preventive action was taken. In the case when the action has been taken and the event occurs, the loss that is prevented is a part of the overall loss L. It is not necessary that event will occur and the cost is justified but the overall benefit should be for the long term.

As discussed in Roulin (2007), the PEV of a forecast system can be simply defined by:

(i) Precipitation (>15mm/day)



(ii) Precipitation (>65mm/day)



◀Figure 10

Reliability and sharpness diagrams for day 1 (a, d), day 2 (b, e) and day 3 (c, f) for precipitation forecasts of NEPS-R (a, b, c) and NEPS-G (d, e, f). 24 h accumulated Precipitation (i) > 15 mm/day and (ii) > 65 mm/day are considered as events

$$PEV = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}}$$

where $E_{climate}$, $E_{forecast}$, and $E_{perfect}$ are the possible expenses that one can bear while taking preventive action based on the information related to the climatology, forecast methods, and near perfect observation, respectively. Expenses in the case of the perfect forecast ($E_{perfect}$) will be the least among all the three expenses. In the forecasting system when $E_{forecast} < E_{climate}$, decisions can be made using the information based on the forecasting system. Using the forecast probabilities and economically beneficial thresholds, decision makers can achieve action recommendations (Lopez et al., 2020; Fundel et al., 2019). The perfect score of economic value is 1. The user will benefit from the forecast when $PEV > 0$.

We assessed the value of the precipitation forecast from the NEPS-R prediction system using a cost/loss decision model and compared it with the corresponding results from NEPS-G. Simulated precipitation has been assessed for rainfall amounts exceeding 15 mm/day and 65 mm/day. Figure 11i shows that NEPS-G provides a slightly higher peak value than NEPS-R in the case of rainfall amount exceeding 15 mm/day. However, it is clear from this figure that varied users with different cost/loss ratios can exploit the positive value of the NEPS-R forecast by choosing different probability thresholds for decision making. In the case of heavy precipitation (> 65 mm/day) the peak value occurs at a very small value of α ($= C/L$) because the peak value occurs at $\alpha = \bar{o}$ (Richardson, 2000) where \bar{o} is the climatological mean frequency of the event. Since heavy precipitation is a rare event the peak value occurs at a very low value of α . It can be noticed from Fig. 11i that though the maximum value obtained from NEPS-G forecast for moderate precipitation amount (> 15 mm/day) is marginally higher, heavy precipitation (> 65 mm/day) forecast in NEPS-R (Fig. 11ii)

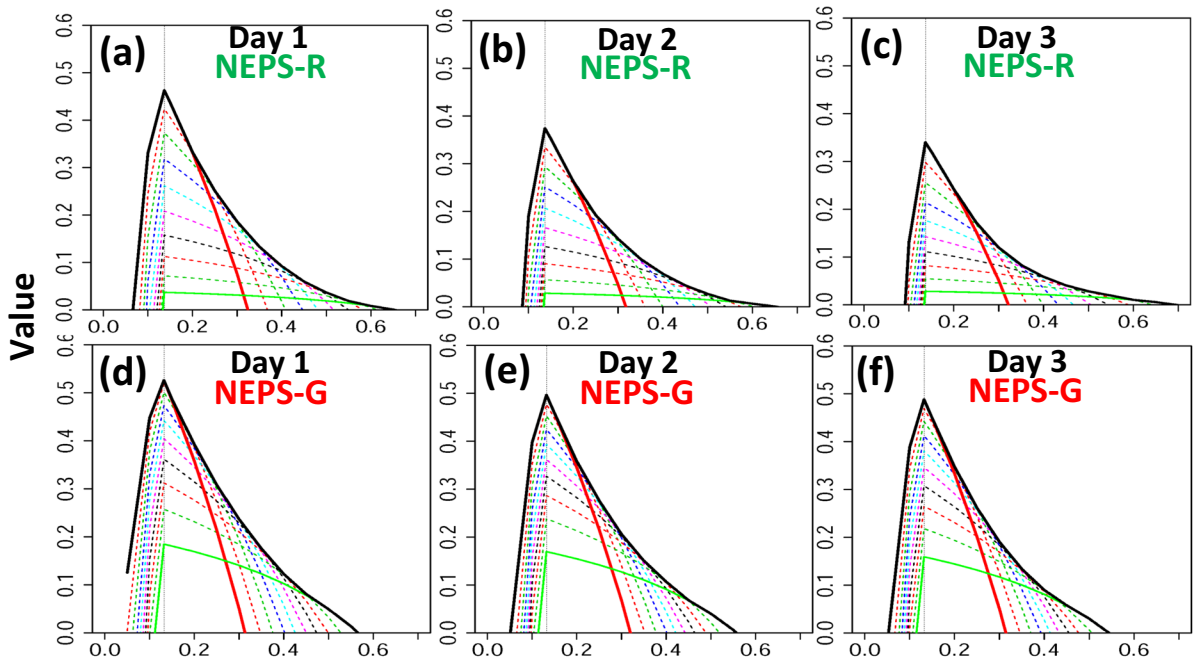
provides a higher maximum value compared to the NEPS-G in all three forecast lead days. The range of users getting positive values for heavy precipitation from NEPS-R and NEPS-G forecasts remains nearly the same.

5. Diurnal Variation

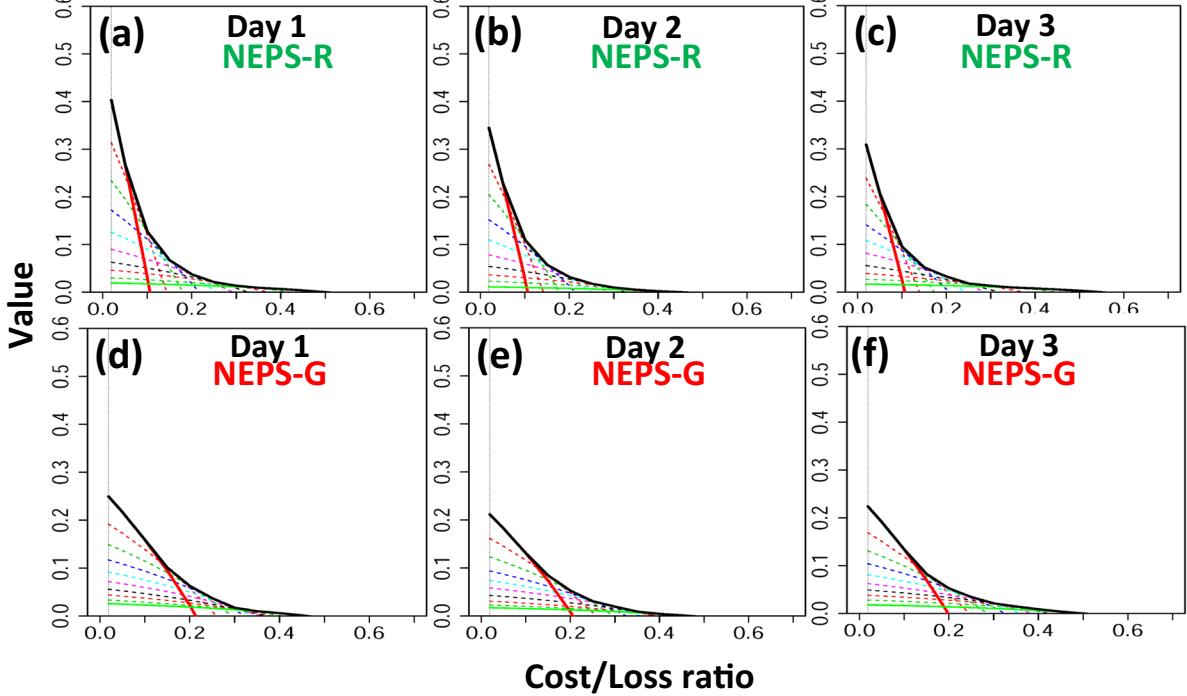
Capturing diurnal variations in precipitation poses a complex challenge particularly over the tropics, and the ability of numerical models to depict these patterns can vary region to region. High resolution regional models, by explicitly incorporating convection and detailed local topography, become pivotal in simulating orographic effects that impact diurnal precipitation patterns. Many studies show an improvement in representing the diurnal cycle of precipitation at a convection permitting scale compared to their coarser resolution version (Clark et al., 2007; Lean et al., 2008; Mamgain et al., 2018a).

The evaluation of diurnal variations in rainfall has been conducted for both the NEPS-R domain and the core monsoon domain as shown in Fig. 1. The core monsoon zone of India is crucial for understanding the dynamics of the Indian summer monsoon (Mandke et al., 2007; Rajeevan et al., 2010). This region encompasses the continental tropical convergence zone, which typically fluctuates during the peak monsoon months, serving as a representative indicator of the intensity of Indian summer monsoon rainfall. For our analysis, we focus on the area between latitudes 18°N and 28°N, and longitudes 73°E and 82°E, falling within the core monsoon zone and considering only the Indian land region. This specific area has also been used to calculate diurnal variations in precipitation. In this section, the diurnal cycle of precipitation during 1st August to 30th September 2019 from both global and regional models has been compared with respect to the observation for the whole of the regional model domain (Fig. 12a) as well as for the monsoon core zone (Fig. 12b). We have used GPM IMERG data and those were further re-gridded to the coarser resolution model for a fairer evaluation. Those data are available at hourly intervals.

(i) Precipitation (>15mm/day)



(ii) Precipitation (>65mm/day)



◀Figure 11

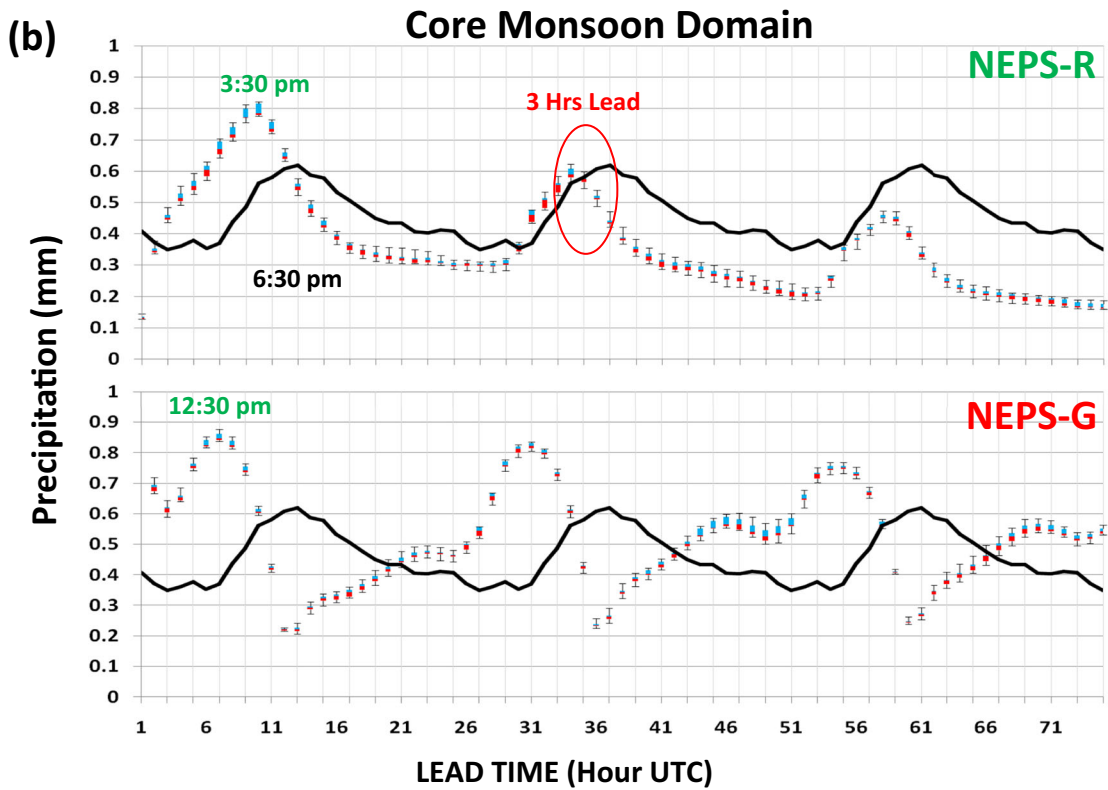
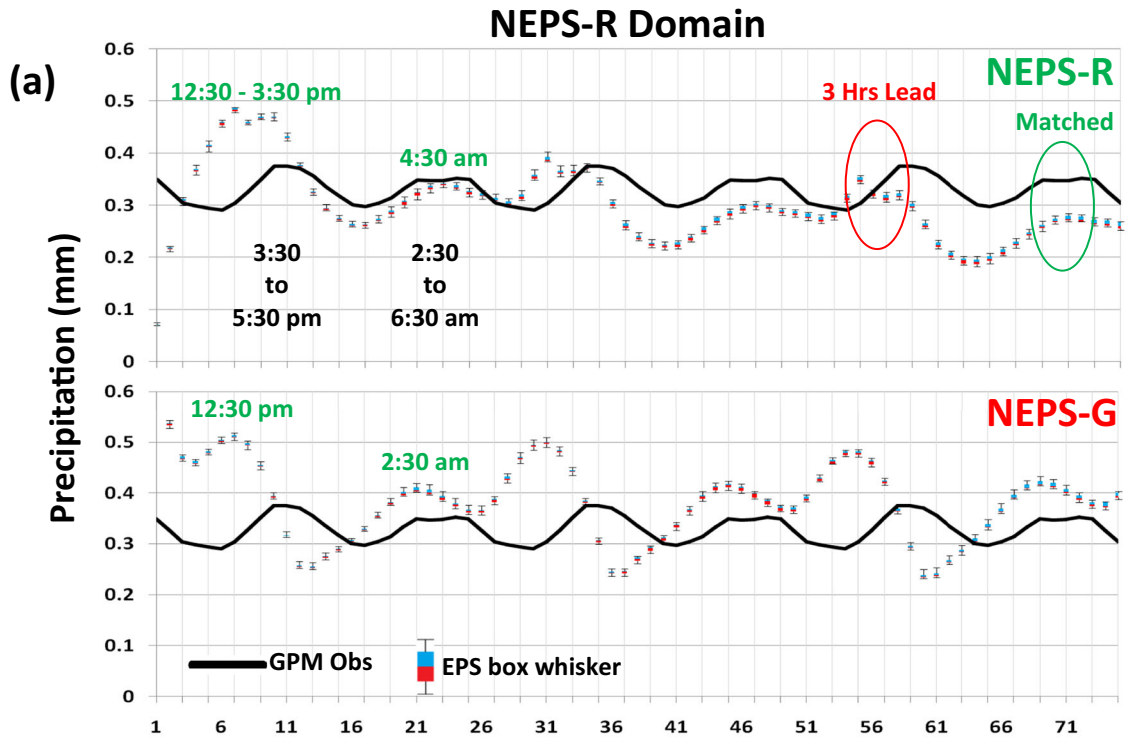
Variation of Economic Value of precipitation greater than (i) 15 mm/day and (ii) 65 mm/day for day 1 (a, d), day 2 (b, e) and day 3 (c, f) forecasts from NEPS-R (a, b, c) and NEPS-G (d, e, f) with cost/loss ratio. Different curves correspond to P_t ranging from 0.09 indicated with solid red line to 0.91 indicated with solid green line. Rest of the P_s (0.18, 0.27, ..., 0.72, 0.82) are depicted with dashed lines. The black curve envelope indicates overall value added by all threshold values of the probability forecasted

The precipitation amount in NEPS-R during the first two hours is very low and unrealistic in both domains (Fig. 12a and b). We are not using 4-km analysis from convective-scale data assimilation in NEPS-R as this is the area of ongoing research and that could be a primary cause for the spin-up of rainfall details in the forecast. In ‘All Indian’ domain (Fig. 12a), the frequency of variation in precipitation is higher than that in the monsoon core zone (Fig. 12b). For India as a whole, different frequencies of variations in precipitation from different regions are combined to form the final wave which is expected to be much smoother than any region-specific variation in precipitation. During the first 11 h (except for the initial 3 h, those needed to get the spin-up details), NEPS-R tends to over-predict the intensity and this wet bias reduced in subsequent hours as shown in Fig. 12a. After nearly 34 h, it further reduces to dry bias. In NEPS-R, this negative trend in precipitation with lead time is probably derived from NEPS-G lateral boundary conditions, as a similar pattern is also noticed in the global model. According to the observations, there are two occasions when peaks in precipitation may occur; those are between 3:30 to 5:30 pm (late afternoon) and 2:30 to 6:30 am (early morning). Early morning maxima in precipitation of NEPS-R is well in phase with the observations but there is a lead of nearly 3 h in late afternoon precipitation that is between 12:30 to 3:30 pm. NEPS-G also exhibits two diurnal peaks for ‘All India’ domain like those in the observations but the forecast and the observed peaks are not in phase. The late afternoon precipitation has a higher intensity than the early morning precipitation in both models. We can notice here that the overall precipitation amount and mainly the peaks are higher in the case of NEPS-G as compared to NEPS-R. That can

be attributed mainly to the fact that the majority of the events in NEPS-G are from light rainfall episodes and those cases are over-predicted due to parameterization in convection and hence contributed to the total rainfall amount. For the monsoon core domain shown in Fig. 12b, a decrease in precipitation intensity with lead time can also be noticed in both models. There is one diurnal peak in the evening around 6:30 pm and NEPS-R is simulating it nearly 3 h earlier than that, at 3:30 pm. NEPS-G is indicating two different peaks and the phase does not agree with that of the observations. The higher diurnal peak in NEPS-G is nearly 6 h earlier than the observation timings. Overall, NEPS-R has demonstrated the better ability to capture the diurnal cycle and the intensity of convective precipitation compared to the global NEPS-G, bringing it closer to observed values. Further enhancement of this forecast can be achieved through the application of appropriate bias correction methods, a topic which is beyond the scope of the present study.

6. Summary and Conclusions

The intention of this paper is to discuss the skill of the operational regional EPS with respect to its driving global model and how the regional model can add any value to the short-range forecast of global EPS operational in NCMRWF. For the comparison of the performances of the two forecasting systems, we have used 11 members of both global and regional models with the initial condition at 00 UTC only. Considering a few verification scores may give a different impression due to small displacements or a temporal shift of the forecast. The forecast based on grid-scale only will not provide the best usable information. So, a comparison has been done using a set of forecast–observation pairs with the help of recommended EPS verification metrics and also including a neighbourhood approach that can determine scales where the model has the desired skill. We also computed the potential economic value of the forecast and finally, diurnal variations in precipitation are analyzed. In this comparative study, the following main points have been noticed.



◀Figure 12

Diurnal variation in area averaged precipitation (mm) calculated for **a** All India domain that is 62° E-106° E; 6° S-41° N, and **b** Monsoon core domain that is 73–82° E; 18–28° N, as simulated by NEPS-R and NEPS-G. Time shown in IST is equivalent to + 05:30 UTC

- Enough ensemble spread is required to represent all the features of the distribution of possible outcomes. We are using 11 members and those are already under sampled. Here both models show under-dispersive nature in U850 and precipitation. Rank histogram statistics indicated wet bias in the global EPS. That is plausibly due to an overestimated count of light precipitation events. Similarly, dry bias is noticed in regional EPS that is increasing with lead time.
- Spread and RMSE in the case of U850 are nearly consistent in both models. The spread in precipitation is much improved in the case of the regional EPS compared to the global EPS. RMSE in U850 and precipitation are however slightly better in global EPS. That can be attributed to the fact that the error computation was done with respect to observation/analysis data at 25 km coarse grid resolution that could generally favor forecast at nearer grid size, here that is 12 km grid of the global EPS compared to the regional EPS at 4 km grid size.
- Skill scores indicate that both models have reliability, resolution, and a tendency to discriminate between the events. However, global EPS skill scores are higher in the case of U850. In the case of precipitation, global EPS scores are also better in the statistics of events including moderate precipitation (> 15 mm/day). For heavy precipitation (> 65 mm/day), the area skill score and rank probability scores of regional EPS are higher than that of global EPS. There are limitations of the grid point-based spatial measurement of errors that arises from the 'double penalty problem' where the error associated with a weather system displaced in space is counted worse than the cases of complete miss or a false alarm.
- Fractions skill scores of precipitation greater than 15 mm/day in the day 1 forecast of regional EPS

are better than those of global EPS. Day 2 and day 3 forecasts of global EPS is better for smaller grid distance, but regional EPS scored higher for increased grid distance. Regional EPS outperformed the global EPS for precipitation greater than 65 mm/day.

- The reliability diagram of U850 in global EPS has more reliable curves compared to the regional EPS. Precipitation greater than 15 mm/day as simulated in regional EPS is more reliable than those in global EPS in all the forecast days mainly for higher forecast probability (> 20%). For the high precipitation events (> 65 mm/day), regional EPS is scoring above the global EPS at the highest forecast probability which is near 90%.
- For precipitation exceeding 15 mm/day, global EPS provides a slightly higher peak of potential economic value than the regional EPS. However, regional EPS provides positive values for a larger range of cost/loss ratios of the forecast for different probability thresholds. For heavy precipitation greater than 65 mm/day, regional EPS provides a higher maximum value compared to the global EPS in all the three forecast lead time days.
- For the spatial precipitation averaged over the 'All India' domain and the 'monsoon core' domain, regional EPS over-predicts the intensity during the 4th to 11th hours of the precipitation forecast and after that at nearly the 34th hour, it further reduces to dry bias. The peaks in precipitation are also over predicted in global EPS but the decrease in precipitation intensity with lead time forecast is at a slower rate. For 'All India' domain, the early morning maxima in precipitation of regional EPS are well in phase with the observations but there is a lead of nearly 3 h in the late afternoon precipitation forecast. Global EPS exhibits these two diurnal peaks for 'All India' domain as those in the observations but overall, the precipitation forecast and the observed peaks are not in phase. Late afternoon precipitation maxima have a higher intensity than the early morning maximum precipitation in both models. Overall precipitation amount and mainly the peaks are higher in the case of global EPS as compared to the regional EPS.

This study suggests the added benefit of using a convection-permitting model for different applications of ensemble forecasting. NEPS-R is expected to provide more information related to the extreme precipitation and diurnal variation. On the other hand, NEPS-G can provide the large-scale patterns and general features more effectively. Still, there are challenges to assess the location specific probabilistic forecasts. The numerical models at the near convection scale may not increase the forecast accuracy as the small-scale errors can also increase because day-to-day random errors are not being included. A major issue in the case of verification and identifying systematic biases of surface variables is the limitations of high-resolution quality-controlled observation.

The uncertainty in the verifying analysis or the initial conditions of the ensemble members is a critical factor in determining the accuracy of the final forecast and making decisions based on them. Any future developments in NEPS-R could include a strategy to perturb initial conditions, considering small-scale uncertainties. Sensitivity experiments can be conducted to fine-tune the model, progressing towards perturbations in model physics. Regional models with explicit convection is expected to have some predictability up to 3 days whereas beyond that period large scale features plays important role. Uncertainty in weather prediction can be addressed by increasing the ensemble size in a single ensemble prediction system, as well as by using different methods such as multi-model ensembles and bias correction techniques. Further, predicting weather using Artificial Intelligence or Machine Learning (AI/ML) models has become a new and emerging field. There are many new research studies where AI/ML models have shown remarkable results in terms of predicting heavy rainfall events over India (Narang et al., 2024; Pham et al., 2020; Subrahmanyam et al., 2021). The European Centre for Medium Range Weather Forecasting (ECMWF) has launched the AI based integrated forecasting system (AIFS) as part of its experiment suite. However, the physics-based numerical weather prediction models still hold importance. They are required for the data assimilation processes to provide initial conditions as well as to create high resolution and accurate reanalysis datasets which will then be used for

training the ML models. Therefore, the requirement of improving the physics-based models is very much relevant and must go hand-in-hand with the advances in the AI/ML based models.

Acknowledgements

The authors gratefully acknowledge the Ministry of Earth Sciences (MoES; India), for providing the resources and encouragement required for this study. We thank the anonymous reviewers for insightful comments and suggestions on this work. The GPM-IMERG data were provided by the NASA Goddard Space Flight Center's Precipitation Measurement Missions Science Team and Precipitation Processing System, which develops and computes GPM-IMERG as a contribution to GPM, and archived at the NASA GES DISC.

Author contributions Ashu Mamgain Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft S Kiran Prasad Data curation, Software Abhijit Sarkar Methodology, Writing – review & editing Gauri Shanker Data curation, Software Anumeha Dube Writing – review & editing Ashis Mitra Project administration, Resources

Funding

The authors have not disclosed any funding.

Data availability

The datasets generated and analyzed during the current study are not publicly available but are available from the MoES, India on reasonable request.

Declarations

Conflict of Interest None.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

REFERENCES

- Anderson, J. S. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integration. *Journal of Climate*, 9(7), 1518–1530. [https://doi.org/10.1175/1520-0442\(1996\)009%3c1518:AMFPAE%3e2.0.CO;2](https://doi.org/10.1175/1520-0442(1996)009%3c1518:AMFPAE%3e2.0.CO;2)
- Bishop, C. H., Etherton, B. J., & Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform kalman filter. part1: Theoretical aspects. *Monthly Weather Review*, 129, 420–436. [https://doi.org/10.1175/1520-0493\(2001\)129%3c0420:ASWTET%3e2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129%3c0420:ASWTET%3e2.0.CO;2)
- Bowler, N. E. (2008). Accounting for the effect of observation errors on verification of MOGREPS. *Meteorological Applications*, 15, 199–205. <https://doi.org/10.1002/met.64>
- Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., & Beare, S. E. (2008). The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134(632), 703–722. <https://doi.org/10.1002/qj.234>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078%3c0001:VOFEIT%3e2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078%3c0001:VOFEIT%3e2.0.CO;2)
- Cafaro, C., & Coauthors.. (2021). Do Convection-permitting ensembles lead to more skillful short-range probabilistic rainfall forecasts over tropical East Africa? *Wea. Forecasting*, 36, 697–716. <https://doi.org/10.1175/WAF-D-20-0172.1>
- Candille, G., & Talagrand, O. (2008). Impact of observational error on the validation of ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, 134, 959–971. <https://doi.org/10.1002/qj.268>
- Capecchi, V. (2021). Reforecasting two heavy-precipitation events with three convection-permitting ensembles. *Weather Forecasting*, 36, 769–790. <https://doi.org/10.1175/WAF-D-20-0130.1>
- Clark, A. J. (2012). An overview of the 2010 hazardous weather testbed experimental forecast program spring experiment. *Bulletin of the American Meteorological Society*, 93(1), 55–74.
- Clark, A. J., Gallus, W. A., & Chen, T. C. (2007). Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Monthly Weather Review*, 135(10), 3456–3473. <https://doi.org/10.1175/Mwr3467.1>
- Clayton, A. M., Lorenc, A. C., & Barker, D. M. (2013). Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Quarterly Journal of the Royal Meteorological Society*, 139(675), 1445–1461. <https://doi.org/10.1002/qj.2054>
- Duc, L., Saito, K., & Seko, H. (2013). Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus A*, 65, 18171. <https://doi.org/10.3402/tellusa.v65i0.18171>
- Ferrett, S., Frame, T. H. A., Methven, J., Holloway, C. E., Webster, S., Stein, T. H. M., & Cafaro, C. (2021). Evaluating convection-permitting ensemble forecasts of precipitation over Southeast Asia. *Weather Forecasting*, 36, 1199–1217. <https://doi.org/10.1175/WAF-D-20-0216.1>
- Frogner, I.-L., Singleton, A. T., Køltzow, M. Ø., & Andrae, U. (2019). Convective-permitting ensembles: Challenges related to their design and use. *Quarterly Journal of the Royal Meteorological Society*, 145(S1), 90–106. <https://doi.org/10.1002/qj.3525>
- Fundel, V. J., Fleischhut, N., Herzog, S. M., Göber, M., & Hagedorn, R. (2019). Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users. *Quarterly Journal of the Royal Meteorological Society*, 145, 210–231. <https://doi.org/10.1002/qj.3482>
- Golding, B. W., Ballard, S. P., Mylne, K., Roberts, N., Saulter, A., Wilson, C., Agnew, P., Davis, L. S., Trice, J., Jones, C., Simonin, D., Li, Z., Pierce, C., Bennett, A., Weeks, M., & Moseley, S. (2014). Forecasting capabilities for the London 2012 olympics. *Bulletin of the American Meteorological Society*, 95(6), 883–896. <https://doi.org/10.1175/BAMS-D-13-00102.1>
- Gowan, T. M., Steenburgh, W. J., & Schwartz, C. S. (2018). Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, 33(3), 739–765. <https://doi.org/10.1175/WAF-D-17-0144.1>
- Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3), 550–560. [https://doi.org/10.1175/1520-0493\(2001\)129%3c0550:IORHFV%3e2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129%3c0550:IORHFV%3e2.0.CO;2)
- Hersbach, H. (2000). Composition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecasting*, 15(5), 559–570. [https://doi.org/10.1175/1520-0434\(2000\)015%3c0559:DOTCRP%3e2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015%3c0559:DOTCRP%3e2.0.CO;2)
- Hohenegger, C., & Schar, C. (2007). Predictability and error growth dynamics in cloud-resolving models. *Journal of the Atmospheric Sciences*, 64, 4467–4478. <https://doi.org/10.1175/2007JAS2143.1>
- Holloway, C. E., Woolnough, S. J., & Lister, G. M. S. (2012). Precipitation distributions for explicit versus parametrized convection in a large-domain high-resolution tropical case study. *Quart. J. Roy. Meteor. Soc.*, 138, 1692–1708. <https://doi.org/10.1002/qj.1903>
- Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K., & Iguchi, T. (2014). The Global Precipitation measurement mission. *Bulletin of the American Meteorological Society*, 95(5), 701–722.
- Johnson, C., & Bowler, N. (2009). On the reliability and calibration of ensemble forecasts. *Monthly Weather Review*, 137(5), 1717–1720. <https://doi.org/10.1175/2009MWR2715.1>
- Jolliffe, I. T., & Primo, C. (2008). Evaluating rank histograms using decompositions of the chi-square test statistic. *Mon. Wea. Rev.*, 136, 2133–2139. <https://doi.org/10.1175/2007MWR2219.1>
- Kirihiga, S. M., Narasimhan, B., & Balaji, C. (2021). A multi-physics ensemble approach for short-term precipitation forecasts at convective permitting scales based on sensitivity experiments over southern parts of peninsular India. *Journal of Earth System Science*, 130, 68. <https://doi.org/10.1007/s12040-021-01556-8>
- Klasa, C., Arpagaus, M., Walser, A., & Wernli, H. (2018). An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. *Quart. J. Royal Meteor. Soc.*, 144, 744–764. <https://doi.org/10.1002/qj.3245>

- George, G., Lodh, A., Rani, S. I., Mohandas, S., George, J. P., Rajagopal, E. N. (2018). "Implementation of new high resolution NCM analysis-forecast system in Mihir HPCS". NMR/ TR/01/2018: pp. 17.
- Lean, H. W., Clark, P. A., Dixon, M., Roberts, N. M., Fitch, A., Forbes, R., & Halliwell, C. (2008). Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, *136*, 3408–3424. <https://doi.org/10.1175/2008MWR2332.1>
- Lopez, A., Coughlan, E., Bazo, J., Suarez, P., Hurk, B., & Aalst, M. (2020). Bridging forecast verification and humanitarian decisions: A valuation approach for setting up action-oriented early warnings. *Wea. Clim. Extremes*, *27*, 100167. <https://doi.org/10.1016/j.wace.2018.03.006>
- Mangain, A., Sarkar, A., Dube, A., Arulalan, T., Chakraborty, P., George, J. P., Rajagopal, E. N. (2018b). "Implementation of very high resolution (12 km) global ensemble prediction system at NCMRWF and its initial validation". NCMRWF, NMR/ TR/02/ 2018, 25 pp.
- Mangain, A., Rajagopal, E. N., Mitra, A. K., & Webster, S. (2018a). Short-range prediction of monsoon precipitation by NCMRWF regional unified model with explicit convection. *Pure and Applied Geophysics*, *175*, 1197–1218. <https://doi.org/10.1007/s00024-017-1754-0>
- Mangain, A., Sarkar, A., & Rajagopal, E. N. (2019). Medium-range global ensemble prediction system at 12km horizontal resolution and its preliminary validation. *Meteorological Applications*, *27*(1), e1867. <https://doi.org/10.1002/met.1867>
- Mangain, A., Sarkar, A., & Rajagopal, E. N. (2020). Verification of high resolution (12 km), global ensemble prediction system. *Atmospheric Research*, *236*, 104832. <https://doi.org/10.1016/j.atmosres.2019.104832>
- Mandke, S. K., Sahai, A. K., Shinde, M. A., Joseph, S., & Chatopadhyay, R. (2007). Simulated changes in active/break spells during the Indian summer monsoon due to enhanced CO2 concentrations: Assessment from selected coupled atmosphere–ocean global climate models. *International Journal of Climatology*, *27*, 837–859. <https://doi.org/10.1002/joc.1440>
- Mason, I. B. (1982). A model for assessment of weather forecasts. *Australian Meteorological Magazine*, *30*, 291–303.
- Maurer, V., Kalthoff, N., & Gantner, L. (2017). Predictability of convective precipitation for West Africa: Verification of convection-permitting and global ensemble simulations. *Meteorologische Zeitschrift*, *26*, 93–100. <https://doi.org/10.5445/IR/1000059465>
- Mitra, A. K., Momin, I. M., Rajagopal, E. N., Basu, S., Rajeevan, M. N., & Krishnamurti, T. N. (2013). Gridded daily Indian monsoon rainfall for 14 seasons: Merged TRMM and IMD gauge analyzed values. *Journal of Earth System Science*, *122*(5), 1173–1182. <https://doi.org/10.1007/s12040-013-0338-3>
- Mittermaier, M. P., & Csima, G. (2017). "Ensemble versus deterministic performance at the kilometer scale. *Weather and Forecasting*, *32*(5), 1697–1709. <https://doi.org/10.1175/WAF-D-16-0164.1>
- Mittermaier, M. P., Roberts, N., & Thompson, S. A. (2013). A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorological Applications*, *20*(2), 176–186. <https://doi.org/10.1002/met.296>
- Narang, U., Juneja, K., Upadhyaya, P., Salunke, P., Chakraborty, T., Behera, S. K., Mishra, S. K., & Suresh, A. D. (2024). Artificial intelligence predicts normal summer monsoon rainfall for India in 2023. *Scientific Reports*, *14*, 1495. <https://doi.org/10.1038/s41598-023-44284-3>
- Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., & Smith, L. (2006). "Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter*. <https://doi.org/10.21957/ab129056ew>
- Pham, B. T., Le, M. N., Le, T.-T., Bui, K.-T.T., Le Minh, V., Ly, H.-B., & Prakash, I. (2020). "Development of advanced artificial intelligence models for daily rainfall prediction", Atmospheric Research, Volume 237, 2020. *ISSN, 104845*, 0169–8095. <https://doi.org/10.1016/j.atmosres.2020.104845>
- Prasad, S.K., Sarkar, A., Mangain, A. (2019). "Implementation of NCMRWF regional ensemble prediction system (NEPS-R)", NMR/ TR/09/2019
- Rajeevan, M., Gadgil, S., & Bhate, J. (2010). Active and break spells of the Indian summer monsoon. *Journal of Earth System Science*, *119*(3), 229–247. <https://doi.org/10.1007/s12040-010-0019-4>
- Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 649–667. <https://doi.org/10.1002/qj.49712656313>
- Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. <https://doi.org/10.1175/2007MWR2123.1>
- Roulin, E. (2007). Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences Discussions, European Geosciences Union*, *11*(2), 725–737. <https://doi.org/10.5194/hess-11-725-2007>
- Schellander-Gorgas, T., Wang, Y., Meier, F., Weidle, F., Wittmann, C., & Kann, A. (2017). On the forecast skill of a convection-permitting ensemble. *Geosci. Model Dev.*, *10*, 35–56. <https://doi.org/10.5194/gmd-10-35-2017>
- Schwartz, C. S. (2019). Medium-range convection-allowing ensemble forecasts with a variable-resolution global model. *Mon. Wea. Rev.*, *147*, 2997–3023. <https://doi.org/10.1175/MWR-D-18-0452.1>
- Sisodiya, A., Pattnaik, S., & Mohapatra, M. (2022). Localized prediction of rainfall over Odisha using multiple physics ensemble approach. *Journal of Earth System Science*, *131*, 89. <https://doi.org/10.1007/s12040-022-01835-y>
- Subrahmanyam, K. V., Ramsenthil, C., Girach Imran, A., Chakravorty, A., Sreedhar, R., Ezhilrajana, E., Bala Subrahmanyam, D., Ramachandran, R., Kumar, K. K., Rajasekhar, M., & Jha, C. S. (2021). Prediction of heavy rainfall days over a peninsular Indian station using the machine learning algorithms. *Journal of Earth System Science*. <https://doi.org/10.1007/s12040-021-01725-9>
- Tennant, W., & Beare, S. (2014). New schemes to perturb sea-surface temperature and soil moisture content in MOGREPS. *Quarterly Journal of the Royal Meteorological Society*, *140*(681), 1150–1160. <https://doi.org/10.1002/qj.2202>
- Tennant, W. J., Shutts, G. J., Arribas, A., & Thompson, S. A. (2011). Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Wea. Rev.*, *139*, 1190–1206. <https://doi.org/10.1175/2010MWR3430.1>
- Walters, D., & Coauthors. (2017). The met office unified model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 CONFIGURATIONS. *Geoscientific Model Development*, *10*, 1487–1520. <https://doi.org/10.5194/gmd-10-1487-2017>

- Wang, Y., Bellus, M., Wittmann, C., Steinheimer, M., Weidle, F., Kann, A., Ivatek-Šahdan, S., Tian, W., Ma, X., Tascu, S., & Bazile, E. (2011). The Central European limited-area ensemble forecasting system: ALADIN-LAEF. *Quarterly Journal of the Royal Meteorological Society*, 2011(137), 483–502. <https://doi.org/10.1002/qj.751>
- Wastl, C., Wang, Y., Atencia, A., Weidle, F., Wittmann, C., Zingerle, C., & Keresturi, E. (2021). C-LAEF - convection-permitting limited area ensemble forecasting system. *Quarterly Journal of the Royal Meteorological Society*. <https://doi.org/10.1002/qj.3986>
- Weusthoff, T., Ament, F., Arpagaus, A., & Rotach, M. W. (2010). Assessing the benefits of convection-permitting models by neighborhood verification: Examples from MAP D-PHASE. *Monthly Weather Review*, 138(9), 3418–3433. <https://doi.org/10.1175/2010MWR3380.1>
- Wilks, D. S. (2005). *Statistical Methods in the Atmospheric Sciences*. Academic Press.
- Wilks, D. S. (2019). Indices of rank histogram flatness and their sampling properties. *Mon. Wea. Rev.*, 147, 763–769. <https://doi.org/10.1175/MWR-D-18-0369.1>
- Wood, N., Staniforth, A., White, A., Allen, T., Diamantakis, M., Gross, M., Melvin, T., Smith, C., Vosper, S., Zerroukat, M., & Thuburn, J. (2014). An inherently mass-conserving semi-implicit Lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Quarterly Journal of the Royal Meteorological Society*, 140(682), 1505–1520. <https://doi.org/10.1002/qj.2235>
- Yano, J.-I., Ziemiański, M. Z., Cullen, M., Termonia, P., Onvlee, J., Bengtsson, L., Carrasi, A., Davy, R., Deluca, A., Gray, S. L., Homar, V., Köhler, M., Krichak, S., Michaelides, S., Phillips, V. T. J., Soares, P. M. M., & Wyszogrodzki, A. A. (2018). Scientific challenges of convective-scale numerical weather prediction. *Bulletin of the American Meteorological Society*, 99, 699–710. <https://doi.org/10.1175/BAMS-D-17-0125.1>

(Received May 1, 2023, revised May 26, 2024, accepted June 11, 2024, Published online June 25, 2024)