

Verification of operational forecast

1. Purpose

Verification of forecasts issued by RSMC New Delhi is an essential component of the forecast and warning system. It provides an objective assessment of the accuracy, reliability and usefulness of forecasts and helps identify areas for improvement in forecasting techniques, numerical weather prediction (NWP), observational inputs and warning services.

2. Forecast parameters to be verified

RSMC issues and verifies following major parameters:

2.1. Guidance under Severe Weather Guidance under SWFP-South Asia Programme

RSMC New Delhi, as the regional centre in the SWFP-South Asia Programme undertakes systematic verification of the severe-weather guidance provided to participating national meteorological and hydrological services (NMHSs).

Verification is carried out by comparing the RSMC forecast guidance with corresponding observations/analyses for each participating country. The entire region has been divided into 95 met-subdivisions (Fig. 1). The verification is performed separately for each sub-division and different forecast lead times (e.g., Day 1 to Day 5) and hazard categories.

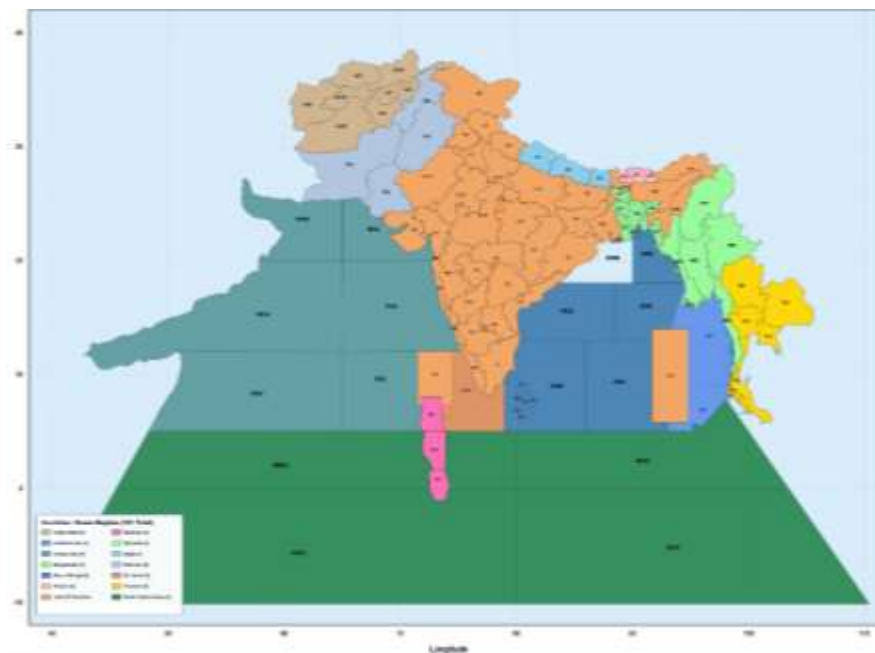


Fig. 1: Geographical domain and sub-regional classification of SWFP South Asia

For categorical forecasts, the following scores are calculated:

- **Probability of Detection (POD):** ability to correctly forecast observed events.
- **False Alarm Ratio (FAR):** proportion of forecast events that did not occur.

- **Critical Success Index (CSI):** overall ability to correctly forecast events, considering both misses and false alarms.
- **Frequency Bias:** indicates whether events are being forecast too frequently or too infrequently.
- **Percentage Correct (PC):** proportion of all forecasts that are correct.
- **Heidke Skill Score (HSS):** skill of the forecast compared with random chance.

The RSMC New Delhi Risk Guidance product provides categorical forecasts rather than fixed numerical forecasts. For example, it indicates Heavy or Very Heavy Rainfall rather than a specific rainfall amount. Therefore, its performance is evaluated using categorical forecast verification methods rather than continuous-error measures such as Mean Absolute Error (MAE) or Root Mean Square Error (RMSE).

The Risk Guidance bulletin classifies rainfall into four categories:

No Heavy Rainfall (NHR: <50 mm/24 h),

Heavy Rainfall (HR: ≥ 50 mm/24 h),

Very Heavy Rainfall (VHR: ≥ 100 mm/24 h), and

Extremely Heavy Rainfall (EHR: ≥ 150 mm/24 h).

Wind is classified as

W1 (<20 kt),

W2 (20–33 kt),

W3 (20–33 kt with an active weather system, or >34 kt without a cyclonic disturbance), and

W4 (>34 kt with an active cyclonic system).

The rainfall thresholds are consistent with commonly used regional warning thresholds, while the higher wind categories consider both wind speed and the associated weather system.

The use of categories is an intentional operational choice. Users generally need to know whether a hazard threshold is likely to be exceeded rather than the precise numerical value. Categorical forecasts also account for the uncertainty of local-scale rainfall and wind forecasts and provide a common risk language for NMHSs across the region.

For verification, the sequence of Risk Guidance bulletins issued by RSMC New Delhi is used. Each bulletin provides forecasts for the next five days at different subdivisions and forecast lead times of +24, +48, +72, +96 and +120 hours. All forecast–observation pairs from the subdivisions and bulletins are combined for calculating the verification statistics.

For rainfall verification, NASA GPM IMERG data at approximately $0.1^\circ \times 0.1^\circ$ resolution and 30-minute intervals are used as the observational reference. Since forecasts are issued at subdivision scale and observations are available at grid points, the observations are aggregated to obtain a representative value for each subdivision. For rainfall, the 48 half-hourly values are first summed to obtain the 24-hour rainfall at each grid point, and the maximum 24-hour rainfall within the subdivision is then selected. This is appropriate because the forecast categories are based on threshold exceedance; using a spatial average could underestimate localized heavy rainfall.

For wind, which is an instantaneous variable, no temporal accumulation is performed. The maximum wind speed across all grid points and observation times within the 24-hour verification period is selected as the representative value. Where wind components are available, scalar wind speed is calculated first. The verification period corresponds to the forecast-valid period, from 0300 UTC to 0300 UTC of the following day. The observed rainfall and wind values are then converted into the same categories used in the forecasts.

Forecast and observed categories are evaluated using exact-match and cumulative approaches. Under the exact-match approach, a forecast is correct only when the forecast and observed categories are identical. Under the cumulative approach, the categories are treated as an ordered hazard sequence: $HR < VHR < EHR$ for rainfall and $W2 < W3/W4$ for wind. A higher forecast category is therefore considered to include the lower hazard thresholds. For example, an EHR forecast with observed HR is considered to have detected the HR threshold, although the higher categories were over-forecast. The baseline categories NHR and W1 are always verified using the exact-match approach.

For each category, forecast and observed results are arranged in a 2×2 contingency table (Table 1) comprising hits, false alarms, misses and correct rejections. Standard verification measures are then calculated.

Table 1: General form of the categorical verification contingency table for a single risk category

	Observed: Event	Observed: No Event
Forecast: Event	Hits (a)	False Alarms (b)
Forecast: No Event	Misses (c)	Correct Rejections (d)

From the contingency table, the following standard verification metrics are computed, each capturing a different dimension of forecast performance:

Table 2: Skill Score for verification of categorical forecast

Metric	Formulation	Scientific Interpretation
Frequency Bias Index (FBI)	$(a + b) / (a + c)$	Ratio of how often a category was forecast to how often it was observed. FBI = 1 is unbiased; FBI > 1 indicates the category is over-forecast (issued more often than it verifies); FBI < 1 indicates under-forecasting. Unlike POD/FAR, FBI says nothing about spatial or temporal correspondence — only about aggregate frequency — so it is best read alongside, not instead of, POD and FAR.
Probability of Detection (POD)	$a / (a + c)$	Fraction of observed events that were correctly forecast. Sensitive only to missed events; ignores false alarms.
False Alarm Ratio (FAR)	$b / (a + b)$	Fraction of forecast events that did not occur. Sensitive only to over-forecasting.
Critical Success Index (CSI)	$a / (a + b + c)$	Overall accuracy for the event class, ignoring correct rejections; appropriate for rare events such as EHR or W4.
Heidke Skill Score (HSS)	$2(ad - bc) / [(a+c)(c+d) + (a+b)(b+d)]$	Accuracy relative to random-chance forecasting, corrected for the base rate of the event; suited to the full four-category classification.
Peirce Skill Score (True Skill Statistic)	$a/(a+c) - b/(b+d)$	Discrimination ability between event and non-event, unaffected by the event base rate; complements HSS.

The Frequency Bias Index (FBI) indicates whether an event is over- or under-forecast, with 1 representing no frequency bias. The Probability of Detection (POD) measures the proportion of observed events correctly forecast, while the False Alarm Ratio (FAR) indicates the proportion of forecast events that did not occur. The Critical Success Index (CSI) considers hits, misses and false alarms and is particularly useful for relatively rare events. The Heidke Skill Score (HSS) measures forecast skill relative to random-chance forecasting.

Verification scores are calculated for different rainfall and wind thresholds, seasons, the complete year, and forecast lead times from +24 to +120 hours. This provides an assessment of how forecast performance varies with season, hazard level and forecast lead time.

The methodology has some limitations. Gridded rainfall products such as IMERG may smooth localized convective maxima, while observational coverage can be limited in mountainous and maritime areas. Therefore, verification results in such regions should be interpreted considering observational uncertainty. In addition, EHR and W4 events are relatively rare, resulting in greater uncertainty in their scores over short periods. Multi-year verification is therefore preferable for assessing performance for the highest-risk categories.

Overall, the methodology compares RSMC New Delhi's categorical Risk Guidance forecasts with observed rainfall and wind, converts the observations into the same hazard categories, and evaluates performance using standard categorical verification measures such as FBI, POD, FAR, CSI and HSS across different seasons and forecast lead times.

The seasonal heavy rainfall (>50 cm) verification scores based on the data during 2024 to August 2026 are presented in Fig. 2 .

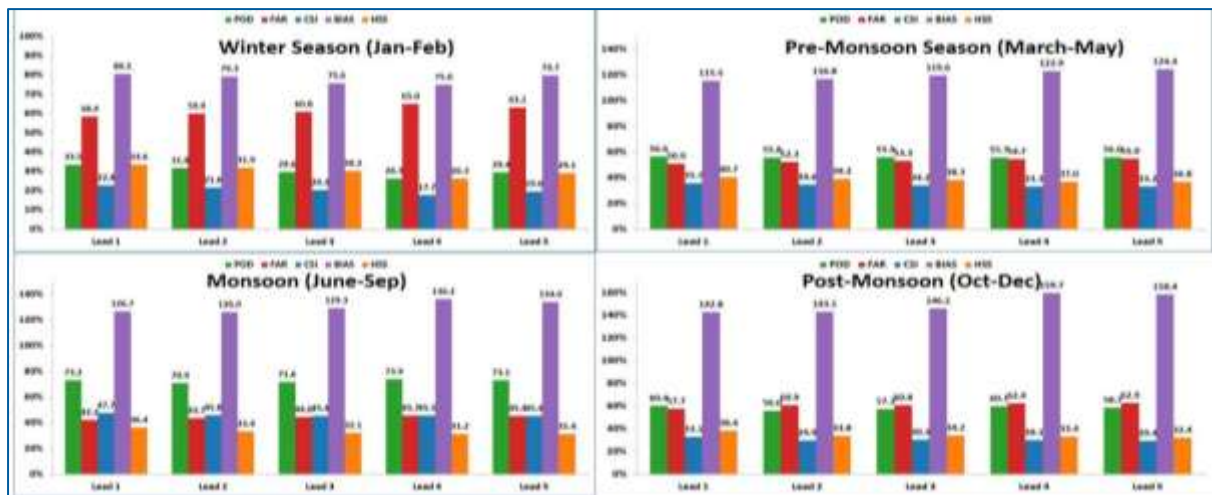


Fig. 2: Seasonal heavy rainfall verification skill scores

Fig. 2 indicates following:

- ❖ Strong seasonal dependence in heavy rainfall (≥ 50 mm) forecast performance.
- ❖ Monsoon season: Highest skill, with POD 73.2%, CSI 48.0%, and relatively lower FAR 42.2%.
- ❖ Winter season: Lowest skill, with POD 33.5% and CSI 22.8%.
- ❖ Pre- and post-monsoon: Moderate detection skill, but with increasing positive frequency bias
- ❖ Frequency bias: Winter: 78% (underforecasting) , Other seasons overforecasting (120-150%)

The skill scores corresponding to different lead periods considering the data during 2024 to August 2026 are presented in Fig.3. It indicates following:

- ❖ Over all, POD is 63-65% for different lead periods
- ❖ False Alarm ratio is 47-51% upto 120 hrs lead period

- ❖ Critical Success Index is 38-41% for different lead periods
- ❖ Bias is mostly positive for various lead periods (over warning)
- ❖ Heidke's Skill Score is 40-42% indicating moderate forecast skill, with the forecasts providing substantially better performance than random chance.

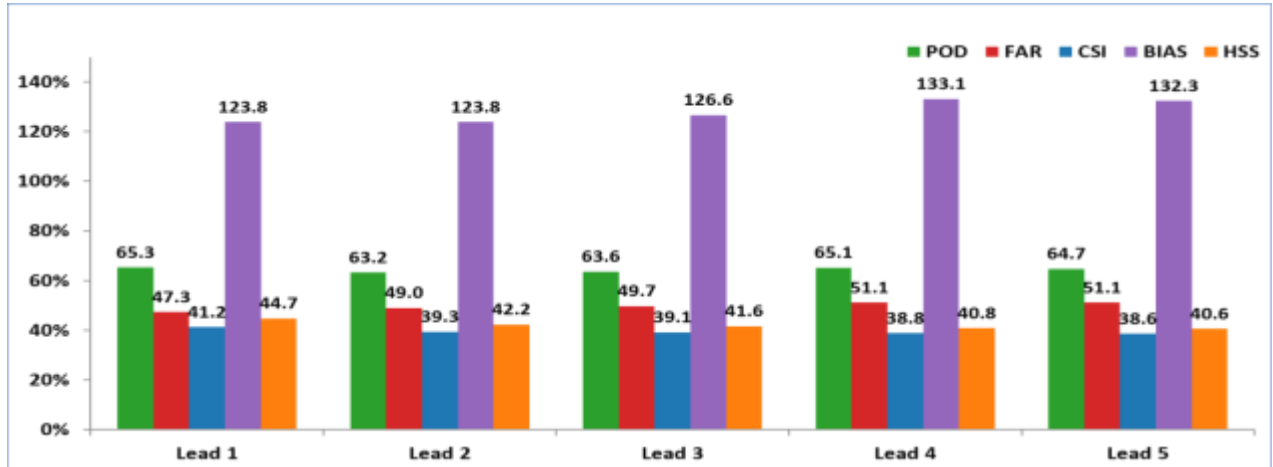


Fig. 3: Heavy rainfall verification skill scores during 2024-26

2.2. Tropical Cyclones

(a) Cyclogenesis

IMD evaluates the performance of operational extended range forecast for cyclogenesis based on the forecasts issued for week 1 and week 2 for AS, BoB and NIO during April 2018 to 2025 for various categories as nil (0% probability), low(1-33% probability), moderate (34-67% probability) and high (68-100% probability). For this purpose, Brier Score Operational (BSO) is calculated and analysed. To evaluate the skill, the climatological probability of cyclogenesis for each week is derived based on the data during 1891-2020 and the same is used to determine reference (BS ref). Subsequently, comparing the operational BSo with reference BS ref, the Brier Skill Score (BSS) has been calculated and analysed (Fig. 4)

The Brier Score Skill (BSS) indicates the degree of improvement of the BS of a forecast (BSO) compared to the BS of a climatological forecast (BSref) as reference, based on an assumed equal occurrence probability of 33% for each event. The forecast is perfect when the BSS is equal to one.

Similarly, in the medium range, IMD commenced the cyclogenesis forecast for next 3 days in 2014, next 5 days in 2018 and next 7 days in 2023 as nil (0% probability), low(1-33% probability), moderate (34-67% probability) and high (68-100% probability) (Fig. 5). To

evaluate the skill of genesis forecast, in a similar manner the BSS is calculated from day1 to day5 for BoB, AS and NIO regions. Every forecast is valid upto 0300 UTC of next day. Details are available [here](#).

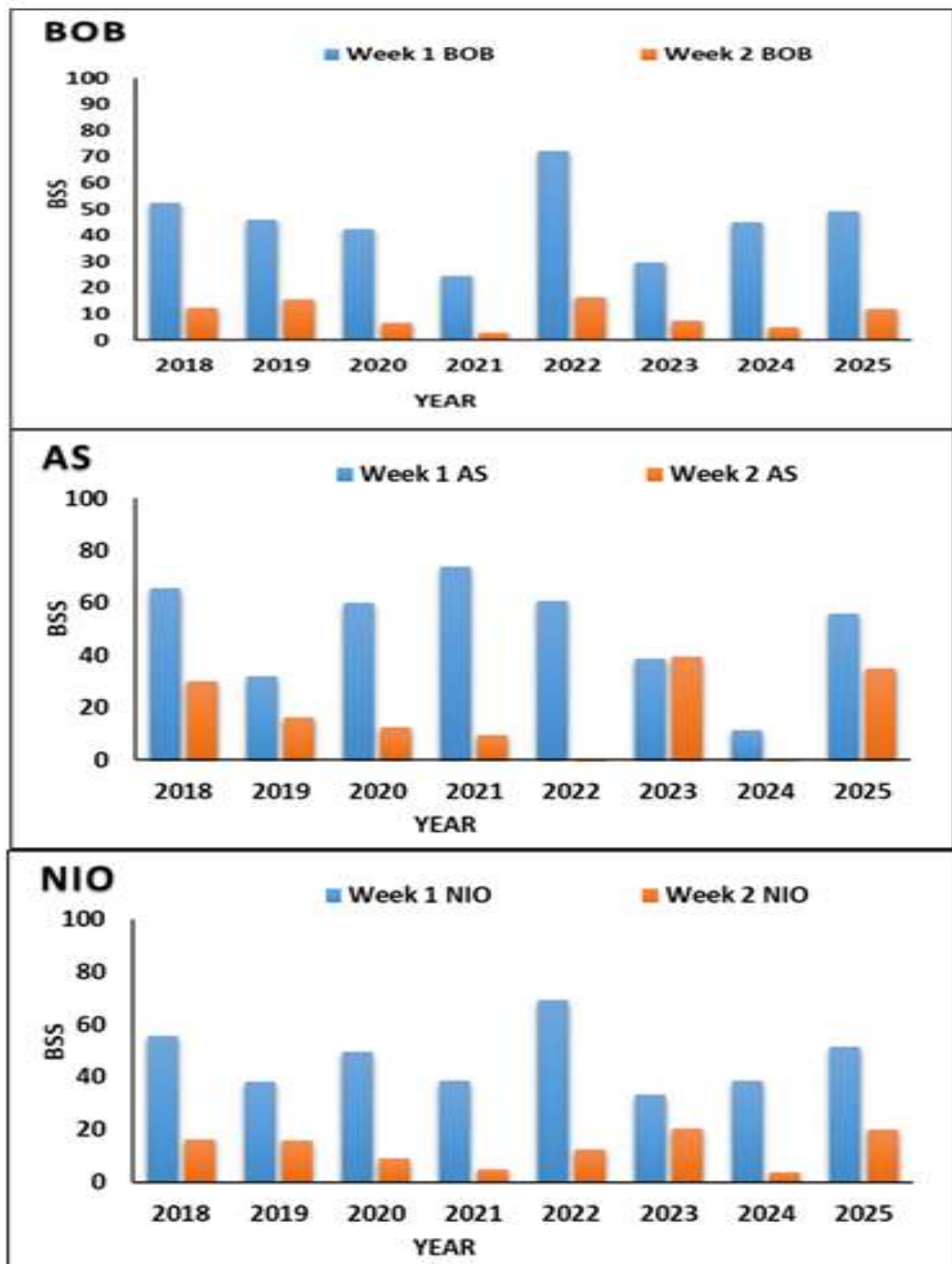


Fig. 4. Annual BSS in the extended range scale for next 2 weeks

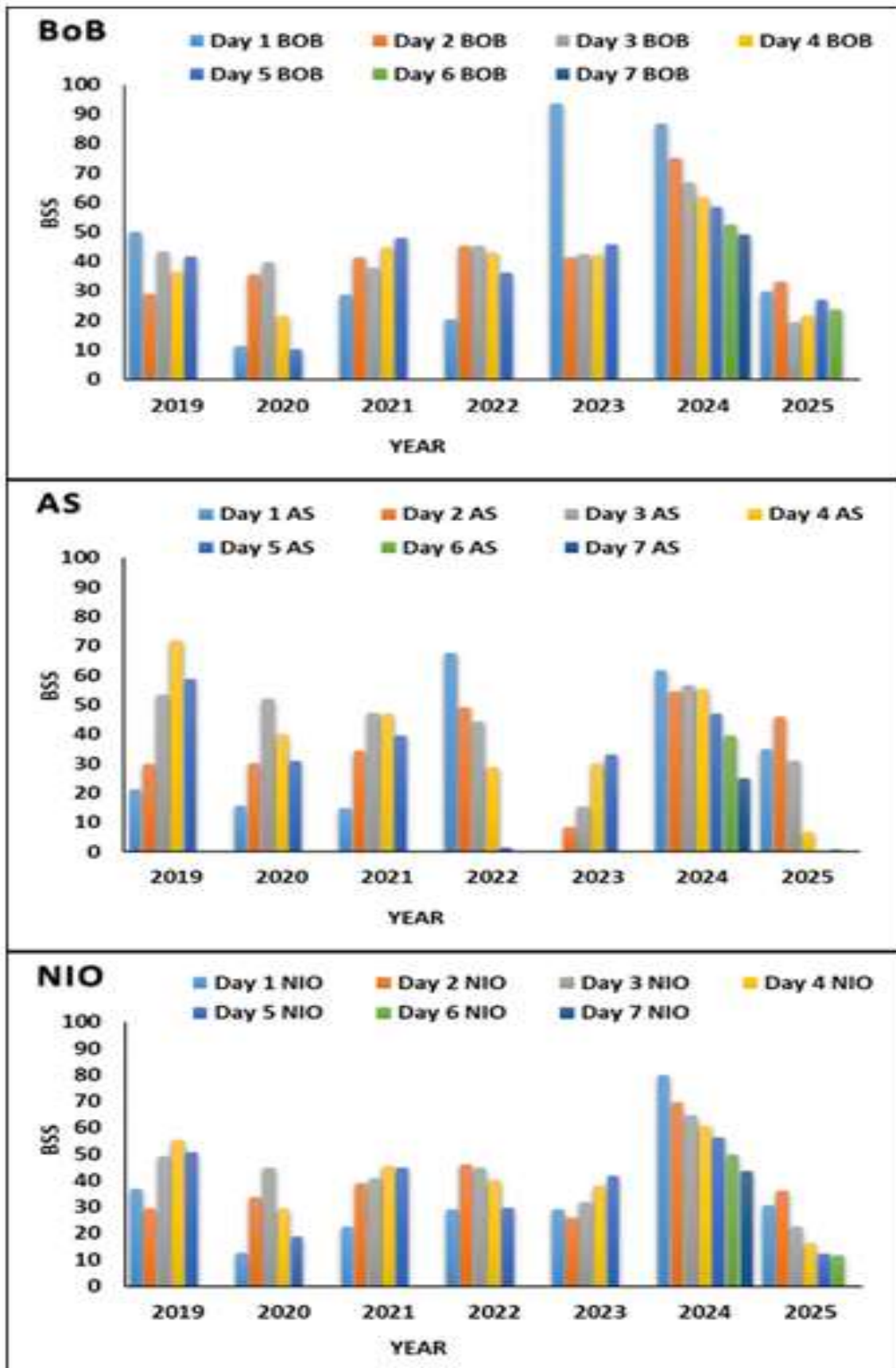


Fig. 5: Annual BSS in the short to medium range scale for next 7 days

(b) Track

To determine the track forecast errors, Direct Position Error (DPE) is calculated as the great-circle distance between the forecast cyclone position and the corresponding position in the finalized best track. The data base includes every six hourly forecasts of track with validity period of 120 hrs. Errors are evaluated at different lead times (12, 24, 36, 48, 72, 84, 96, 108, 120h, etc.) and track forecast skill is assessed relative to a reference climatology/persistence forecast (CLIPER). Further these are compared with previous 5 years errors and skill in track forecast. Details are available [here](#). Annual average track forecast errors and skill during 2025 are presented in Fig.6. Five year comparative track forecast errors during last 5 years (2021-25) and preceding 5 years during (2016-20) are presented in Fig. 7.

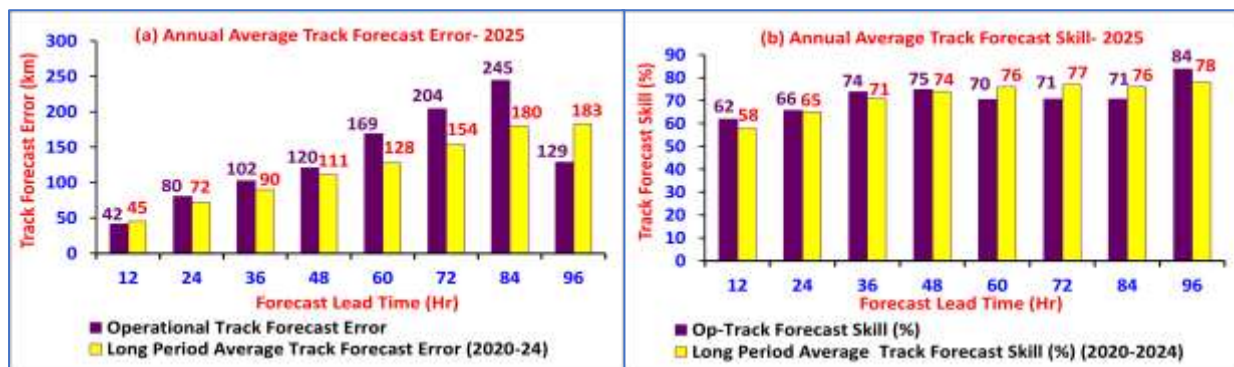


Fig.6 Annual average (a) track forecast error (km) and (b) track forecast skill against the climatology and persistence forecast during 2025 as compared to that during 2020-2024

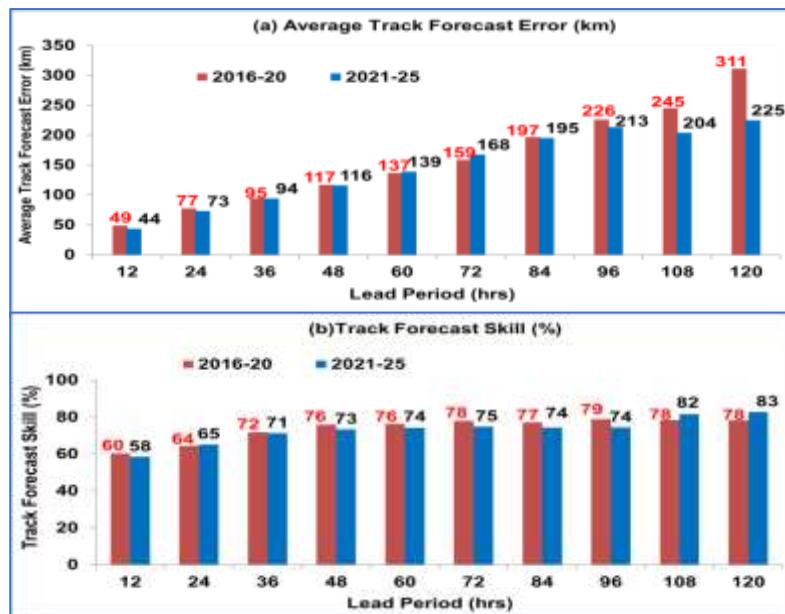


Fig. 7: Comparative average track forecast errors(km) and skill (%) during 2021-25 and 2016-2020

(c) Intensity

Intensity forecast errors are calculated in terms of (a) absolute error (AE) and (ii) root mean square error (RMSE). The data base includes every six hourly forecasts of MSW with validity period of 120 hrs. The average AE is defined as the average of the magnitude of difference between the forecast and actual value of MSW. The mean absolute error (MAE) is calculated for each TC and for all the TCs during the year. The RMSE is also calculated for each TC and for all the TCs during the year, as the RMSE is a measure of dispersion in the errors and indirectly measures the error due to outliers. Higher the number of outliers, more is the RMSE. The skill of the forecast is evaluated against the persistence forecast. The persistence forecast calculates the past 12 hr trend in maximum sustained wind speed (MSW) of the current TC and assumes that TC will intensify/decay with the same trend in MSW during next 120 hrs. Details are available [here](#). Annual average intensity forecast errors and skill during 2025 are presented in Fig.7. Five year comparative intensity forecast errors during last 5 years (2021-25) and preceding 5 years during (2016-20) are presented in Fig. 8.

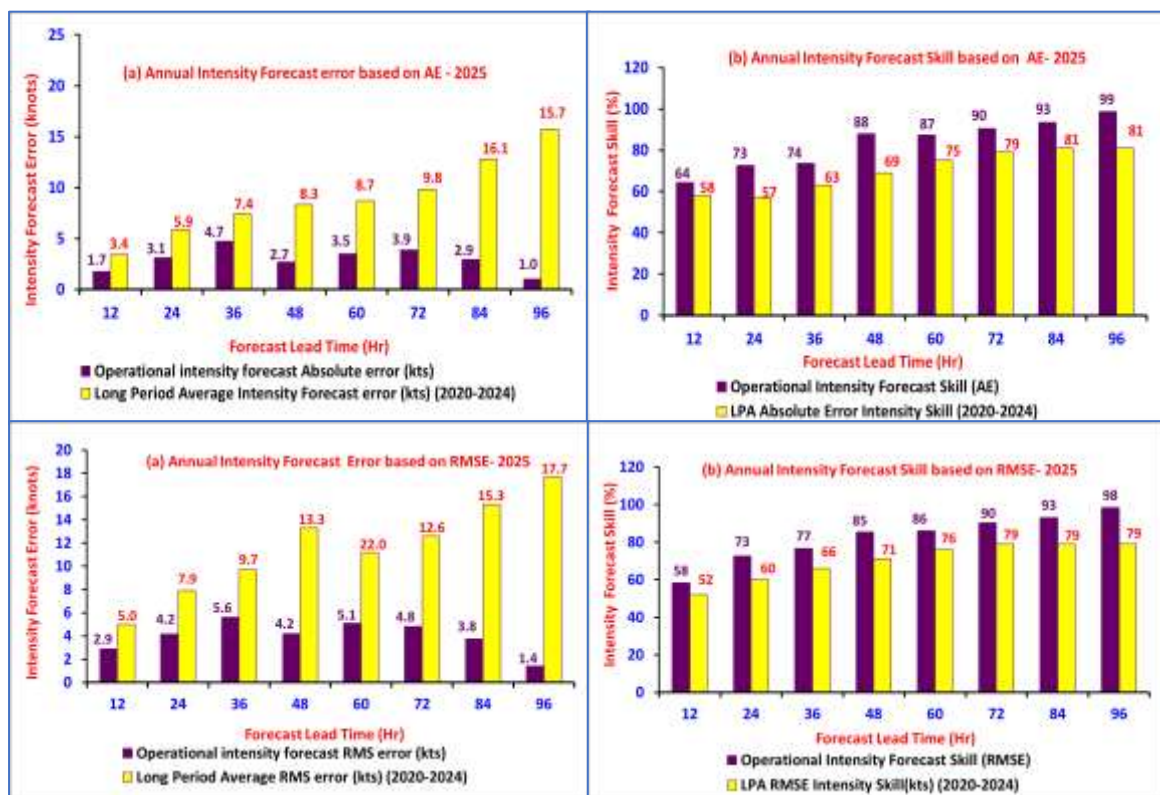


Fig 8: Annual average (a) intensity forecast error (kt) and (b) intensity forecast skill (%) against the persistence forecast during 2025 as compared to that during 2020-2024

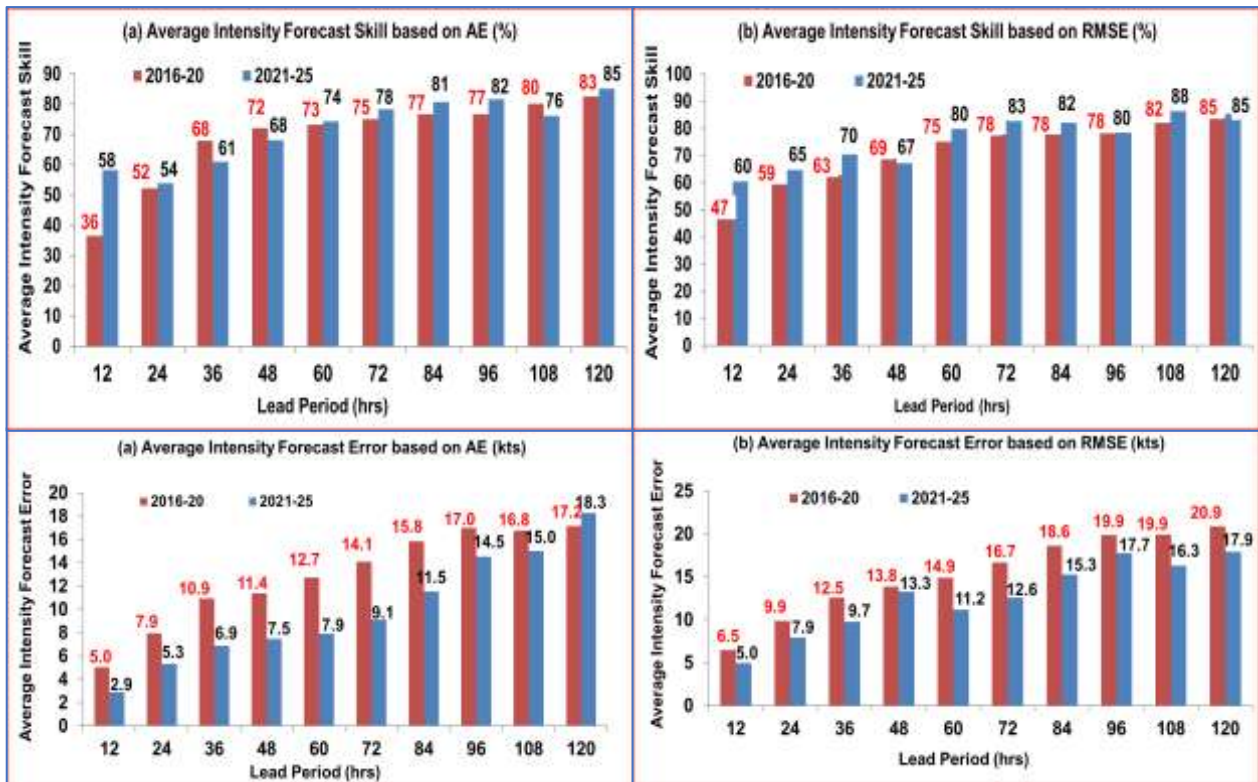


Fig. 9: Comparative average intensity forecast errors (kt) and skill(%) during 2021-25 and 2016-20

(d) Landfall:

The TC landfall forecast errors are analysed by determining the absolute position error and time error. The landfall point forecast error (LPE) is calculated as the distance between the RSMC forecast landfall point and the actual landfall point according to RSMC best-track estimates. Similarly, the landfall time forecast error (LTE) is measured as the difference in the RSMC forecast landfall time and the actual landfall time according to RSMC best-track estimates. Details are available [here](#). Annual average landfall point (km) and time (hrs) forecast errors during 2025 are presented in Fig.9. Five year comparative landfall point (km) and time (hrs) forecast errors during last 5 years (2021-25) and preceding 5 years during (2016-20) are presented in Fig. 10.



Fig 10: Annual average landfall (a) point (km) and (b) landfall time (hrs) forecast errors against the persistence forecast during 2025 as compared to that during 2020-24

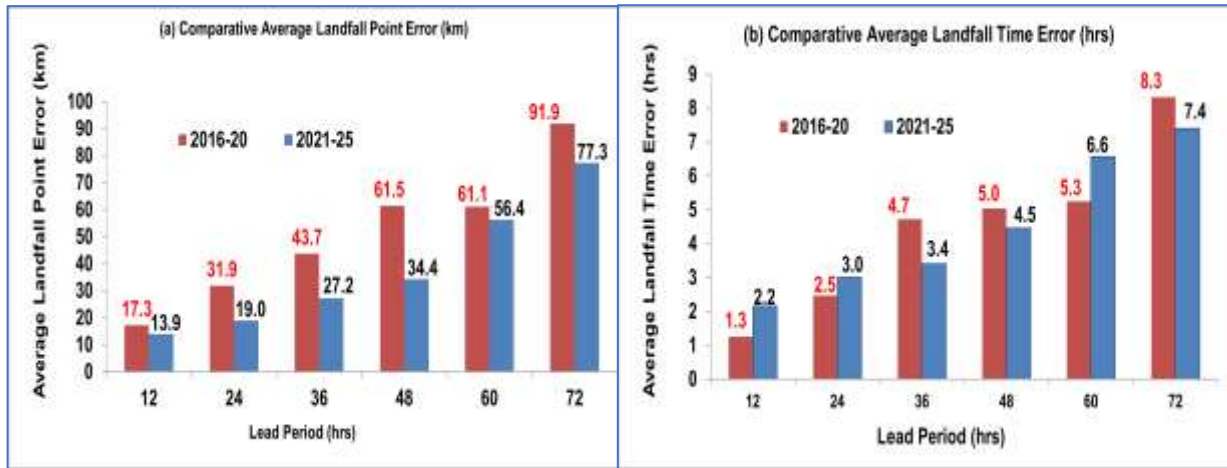


Fig. 11: Comparative average landfall (a) point (km) and (b) time forecast errors (hrs) during 2021-25 and 2016-2020

2.3. Verification of Marine Weather parameters

RSMC carries out verification of 10m maximum sustained wind predicted by different models and MME against ship and buoy observations. Details are available [here](#).

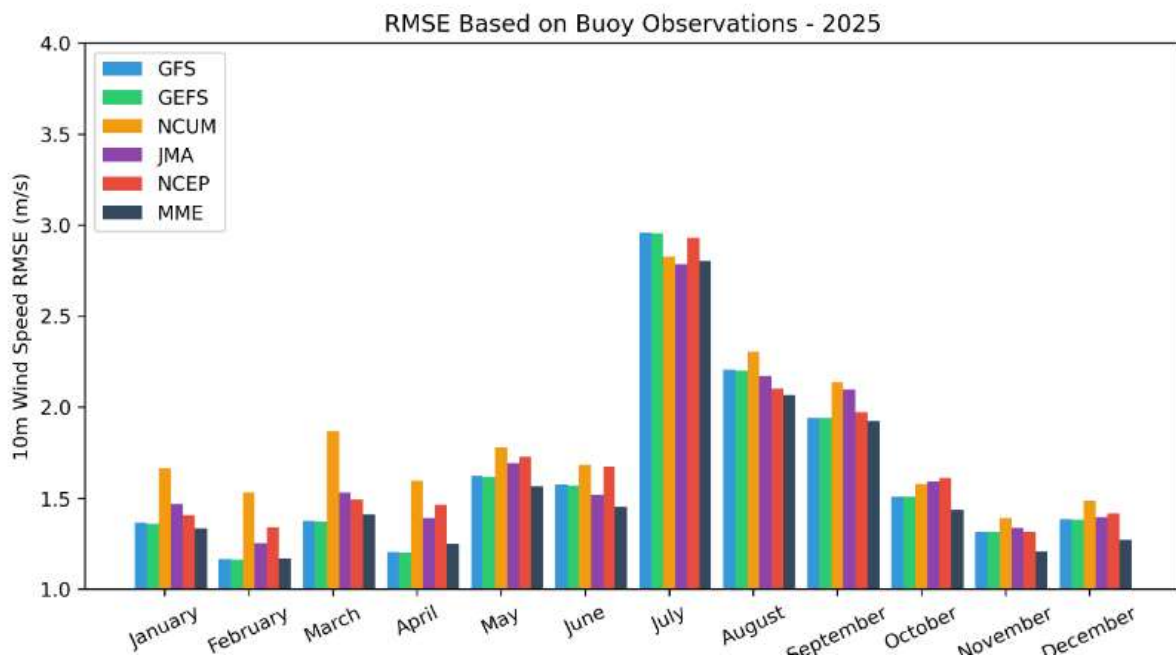


Fig.12: Monthly root mean square error in prediction of 10 m wind by various individual models and multi-model ensemble

The hexbin density plot analysis of 10m wind speed models against observations provides key insights into their performance by visualizing the concentration of data points rather than individual values. The dark black dashed line at 45 degrees represents perfect agreement, serving as a benchmark for model accuracy. High-density regions, shown in red/dark red, indicate where predictions frequently occur, while white and light blue areas represent lower density, helping to reveal overall model behavior.

Clusters along the 45-degree line suggest good agreement between model outputs and observations, whereas deviations above or below indicate systematic biases—with points above implying model underestimation and those below indicating overestimation. The spread of densities provides insight into the variability of model performance, where a compact and dense cluster near the reference line (Black dotted line) signifies higher consistency and accuracy, while wider distributions suggest greater forecast uncertainty.

Comparing different models allows for an evaluation of consistent biases and errors, while assessing the Multi-Model Ensemble (MME) highlights its ability to reduce spread and improve overall agreement with buoy data. The denser concentration near the 45-degree line for MME suggests improved accuracy over individual models.

Furthermore, incorporating a regression line, slope, and R^2 value enhances the analysis by providing quantitative measures of correlation between model predictions and observations. A higher R^2 value for the MME suggests stronger agreement with buoy data. The color gradient (from white to dark red) makes it easier to identify areas where model predictions concentrate, helping to highlight potential biases. A higher density of predictions to the right of the reference line (Black dotted line) confirms a model overestimation bias, while the MME exhibits a reduced bias and a more compact spread.

In summary, the hexbin density plot visually captures agreement, biases, and variability, offering a clear and structured method for evaluating wind speed model performance. The density-based visualization provides a more interpretable and informative alternative to scatter plots, allowing for more effective model assessment and forecasting improvements. Hexbin density plots with verification against buoy & ship are presented in Fig. 13 and Fig. 14.

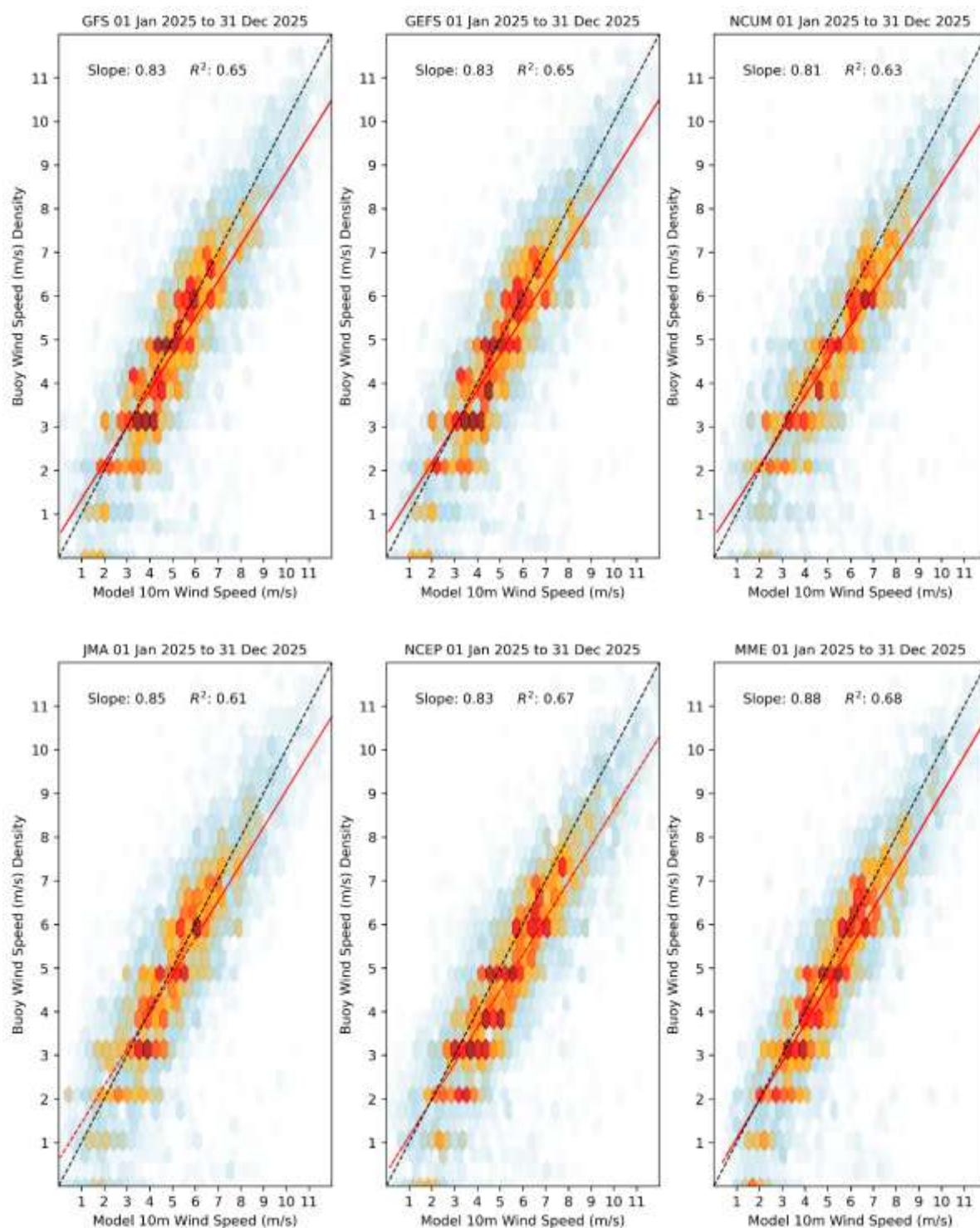


Fig. 13: Hexbin density plots as per verification against buoy data

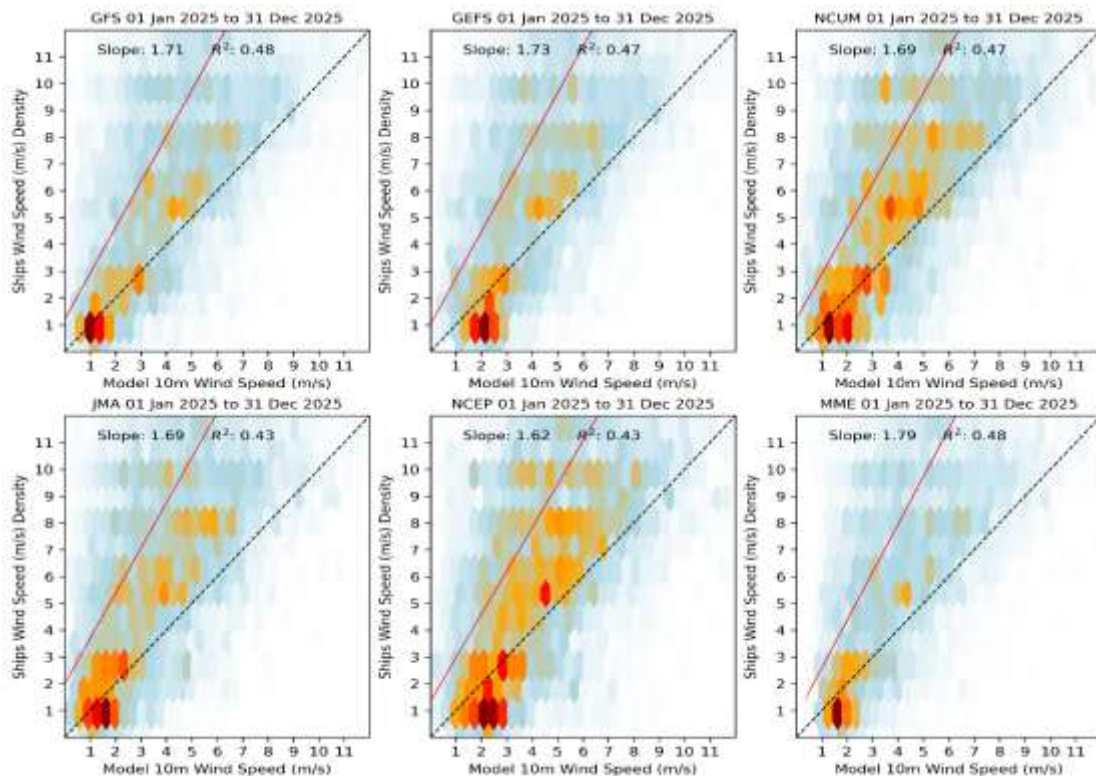


Fig. 14: Hexbin density plots as per verification against ship observations

References:

- Doswell, C. A., R. Davies-Jones, and R. A. McPherson, 1990: On the use of false alarm rates and probability of detection in forecast verification. *Monthly Weather Review*, 118(1), 213–218.
- Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, 15(1), 51–64.
- Gilleland, E., D. A. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5), 1416–1430.
- Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler*, 8, 301–349.
- Huffman, G. J., E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan, 2020: GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06/V07. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD.
- Jolliffe, I. T., and D. B. Stephenson (Eds.), 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd Edition. John Wiley & Sons, Chichester, UK.
- M. Mohapatra, D.P. Nayak, R.P. Sharma, B.K. Bandyopadhyay, 2013, Evaluation of Official Tropical Cyclone Track Forecast over north Indian Ocean Issued by India Meteorological Department, *Journal of Earth System Sciences*, 122, pp 589-601.

- M. Mohapatra, B.K. Bandyopadhyay , D.P. Nayak, 2013, Evaluation of Official Tropical Cyclone Intensity Forecast over north Indian Ocean Issued by India Meteorological Department, Natural Hazards, 68, pp 433-451.
- M. Mohapatra, D.P. Nayak, Monica Sharma, R.P. Sharma, B.K. Bandyopadhyay, 2015, Evaluation of landfall forecast over North Indian Ocean issued by India Meteorological Department, Journal of Earth System Sciences, 124, pp 861-874
- Peirce, C. S., 1884: The numerical measure of the success of predictions. Science, 4(93), 453–454.
- Prakash, S., A. K. Mitra, D. S. Pai, and A. AghaKouchak, 2016: From TRMM to GPM: How well can satellite precipitation products capture extreme precipitation over India? Journal of Hydrometeorology, 17(11), 2825–2841.
- RSMC New Delhi (IMD), 2021: Severe Weather Forecasting Programme (SWFP–South Asia) Regional Operational Manual and Risk Guidance Archive. India Meteorological Department, New Delhi.
- Wilks, D. S., 2019: Statistical Methods in the Atmospheric Sciences, 4th Edition. Academic Press / Elsevier, 840 pp.
- World Meteorological Organization (WMO), 2014: Guidelines on Multi-hazard Impact-based Forecast and Warning Services. WMO-No. 1150, Geneva, Switzerland.
- World Meteorological Organization (WMO), 2020: Guidance on the Verification of Operational Meteorological and Hydrological Warnings. WMO-No. 1229, Geneva, Switzerland.