Environmental Science Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6va00077k

Improving deterministic forecasts of maximum and minimum temperature using machine learning

Harvir Singh, *[ab] Anumeha Dube, [a] Prashant Kumar Srivastava, [b] Raghavendra Ashrit, [a] John P. George [a] and V. S. Prasad [a]

Accurately forecasting near-surface temperature is essential for heatwave and cold-wave warnings and impact-based decision support over India. Deterministic numerical weather prediction (NWP) models show systematic, regionally varying biases that increase with lead times. To improve the reliability of these forecasts, bias correction is essential. This study applies a multivariate machine-learning (ML) bias-correction framework to location specific 2 m maximum (T_{max}) and minimum (T_{min}) temperature forecasts from the operational NWP model at the National Centre for Medium Range Weather Forecasting (NCMRWF). Data from 179 India Meteorological Department (IMD) stations covering the period 2019–2024 were used. Four ML methods, Random Forest (RF), eXtreme Gradient Boosting (XGB), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNNs) were used for bias correction of the forecasts at the 179 stations. The ML models were assessed using continuous metrics like mean error (ME), root mean square error (RMSE), and correlation/Taylor diagnostics. Along with these categorical skills for extremes, metrics like equitable threat score (ETS) and Heidke Skill Score (HSS) (for $T_{max} \geq 30/35$ °C in MAMJ (March–June) and $T_{min} \leq 10/15$ °C in DJF (December–February)), and Relative Economic Value (REV) were used. It is found that ML post-processing substantially reduces bias and error across stations and lead times. For T_{max}, RMSE improvement increases with lead time, typically ~10–15% at Day-1, ~20–30% by Day-5, and frequently >30–40% (locally reaching ~50–60%) by Day-9, especially for XGB/LSTM. For T_{min}, improvements are strongest: XGB improves RMSE by ~25–40% at Day-1, increasing to ~40–60% by Day-7 to Day-9 across many stations. Categorical verification shows consistent improvements in terms of higher ETS/HSS values after bias correction across most stations. Winter T_{min} shows large gains for both thresholds, particularly for $T_{min} \leq 15$ °C. REV analysis indicates that ML-corrected forecasts remain economically useful over a wider range of cost–loss ratios and retain value at longer lead times compared to the raw model. Overall, XGB provides the most consistent improvement across regions and metrics, RF is generally second-best, LSTM shows competitive performance, particularly for T_{max} and at longer lead times, while CNN performs worst. SHAP-based analysis links the corrections to physically meaningful drivers, with T_{max} corrections dominated by boundary-layer/land-surface predictors and T_{min} corrections dominated by radiative and synoptic controls.

Received 12th February 2026
Accepted 25th May 2026

DOI: 10.1039/d6va00077k

rsc.li/esadvances

Environmental significance

India's varied climate—from arid deserts to humid plains and Himalayan cold fronts—makes reliable near-surface temperature forecasts vital for safety and economic stability. Heatwaves during March–June claim lives and disrupt agriculture, while winter cold waves damage rabi crops and trigger health alerts. Operational NWP models at NCMRWF, crucial for IMD warnings, suffer biases such as summer T_{max} overestimation and winter T_{min} underestimation, reducing skill beyond Day-3. Traditional bias-correction methods often miss extremes and regional variability. This study applies advanced ML frameworks (RF, XGB, LSTM, CNN) to station forecasts, improving thresholds, reducing RMSE by up to 60%, enhancing ETS/HSS, and extending reliability—supporting scalable, equitable forecasting across India's IMD network.

1. Introduction

In recent years, particularly over the last decade, the intensity and frequency of extreme temperature-related events, such as heatwaves, have increased significantly. These extremes usually result in substantial socio-economic impacts, including increased mortality, heightened energy demand, and

[a]National Centre for Medium-Range Weather Forecasting, Ministry of Earth Sciences, India. E-mail: harvir@ncmrwf.gov.in; harvir.ncmrwf@gmail.com
[b]Institute of Environmental and Sustainable Development, Banaras Hindu University, India

agricultural losses. In India, heatwaves predominantly affect the northwestern, central, and peninsular regions during the pre-monsoon months of March to June.[1,2] Accurate temperature forecasts are therefore essential for timely warnings, informed decision-making, and effective disaster preparedness and mitigation.

Over the past two decades, significant advances in numerical weather prediction (NWP) systems, driven by improvements in model dynamics, physical parameterizations, data assimilation techniques, and increased computational capacity, have enabled increasingly high-resolution forecasts of near-surface meteorological variables, including temperature. Despite these advancements, NWP models exhibit both systematic biases and random errors in their forecasts arising from various sources, including initial-condition errors, uncertainties in land-surface processes, and physical parameterizations.[3–6] During extreme temperature events such as heatwaves and cold spells, even relatively small forecast biases can substantially affect the timing, intensity, and reliability of warnings, thereby limiting the effectiveness of forecast-based decision support.

To address these limitations, statistical post-processing techniques have been widely applied to improve the quality of temperature forecasts.[7–12] Over India, several studies have demonstrated the effectiveness of statistical bias-correction approaches for operational temperature forecasts.[13–16] They have shown significant improvements in forecast accuracy. While these methods are robust and computationally efficient, they generally correct biases in a single predictand and do not explicitly account for the influence of other environmental or meteorological variables.

1.1 Bias correction using ML

Machine learning (ML) approaches offer a flexible multivariate framework for temperature bias correction by allowing the incorporation of additional predictors, such as humidity, wind, precipitation, and other land-surface variables that influence surface temperature. Because atmospheric variables interact nonlinearly across spatial and temporal scales, correcting a single variable in isolation can introduce inconsistencies. For example, interdependencies among temperature, humidity, and wind mean that adjusting one variable without accounting for the others may disrupt the energy balance and atmospheric dynamics. A multivariable framework is therefore essential to preserve physical coherence while improving accuracy, and the importance of using multiple predictors for bias correction has long been recognized.[17,18]

ML-based methods have the potential to learn complex, non-linear relationships between forecast errors and environmental conditions, thereby further improving forecast accuracy for variables such as temperature, particularly under diverse climatic regimes and extreme conditions. These methods are now well established in the literature and have been successfully applied to temperature bias correction. For example, Niazkar *et al.*[19] corrected the biases in temperature data obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis 5 (ERA5) Land reanalysis at 10 stations in northern Italy using nine ML techniques and found that XGB was the best-performing standalone model, while *K*-nearest-neighbor ensembles achieved the best results at most stations. Guo *et al.*[20] applied bias correction to wind forecasts from the ECMWF Integrated Forecasting System (IFS) Cycle 46r1 and demonstrated that incorporating upper-air variables with low-resolution training data improved surface forecast accuracy while reducing computational effort. Recently, Singh *et al.*[21] applied ML-based bias correction to station-level T_{max} and T_{min} data from the Indian Monsoon Data Assimilation and Analysis (IMDAA) reanalysis over the Indian land region and demonstrated that RF and XGB consistently outperformed other methods in reducing T_{max} and T_{min} biases across most Indian stations. Veldkamp *et al.*[22] in their study demonstrated that by incorporating spatial patterns from NWP grids, CNN produced more skillful probabilistic wind speed forecasts. This showed that spatial feature extraction can add value in post-processing frameworks. However, such spatially aware approaches require gridded training datasets and considerably greater computational resources as compared to station-based post-processing methods.

Although these studies highlight the potential of ML-based multivariable bias correction, comprehensive implementations remain relatively underexplored. Several studies[23–25] suggest that such approaches better capture interactions among atmospheric parameters and lead to improved bias-correction performance.

In recent times there has been an emergence of AI-based weather prediction (AIWP) systems such as Pangu-Weather,[26] GraphCast,[27] and ECMWF's AIFS.[28] These models, for some variables, can show skill comparable with the traditional NWP models. Trotta *et al.*[29] showed that post-processing frameworks originally developed for NWP can be applied directly to AIWP models without modification and can yield comparable accuracy. However, most meteorological centres use traditional NWP systems for operational forecasting. NCUMG is currently the primary operational model for temperature forecasts over India, and correcting its biases through ML methods has value for forecasters.

1.2 Objective of the study

At the National Centre for Medium Range Weather Forecasting (NCMRWF), the operational deterministic forecasting system is the NCMRWF Unified Model Global (NCUMG), which is based on the UK Met Office Unified Model. NCUMG has a horizontal resolution of approximately 12 km with 70 vertical levels and provides forecasts up to 10 days. Further details about the model are presented in Section 2.3. Despite its advanced configuration, systematic biases in near-surface temperature forecasts persist, motivating the need for effective post-processing strategies.

The primary objective of this study is to correct systematic biases in 2 m maximum and minimum temperatures (T_{max} and T_{min}) from NCUMG across 179 stations over India using a multivariate machine learning (ML)-based bias-correction framework. The specific objectives are as follows:

• This study aims to quantify forecast biases in T_{max} and T_{min} at the station level for representative lead times (Day-1 to Day-9), particularly during the summer (MAMJ) and winter (DJF) seasons, and to capture regional and seasonal variability in model errors.

• Apply ML-based bias-correction methods at the station level to generate bias-corrected, location-specific forecasts and enhance the operational usability of NCUMG outputs for temperature monitoring, warnings, and decision support.

The study employs four ML techniques: Random Forest (RF),[30] eXtreme Gradient Boosting (XGBoost),[31–33] Convolutional Neural Networks (CNN),[21,34,35] and Long Short-Term Memory (LSTM).[36,37]

The analysis is based on NCUMG forecasts and corresponding station observations over India (Fig. 1) for the period 1 January 2019 to 31 December 2024, with a lead time of up to 9 days. In addition to temperature, the predictor set includes surface variables (Table 1) such as accumulated precipitation (APCP_24), 10 m wind speed (WS10m), 2 m relative humidity (RH2m), soil temperature (TSOIL1m) 0–0.1 meter below the ground, and other relevant meteorological variables and upper air variables (Table 2), such as geopotential height (HGT500), relative humidity (RH850), air temperature at 850 hPa (T850), and horizontal & vertical wind components at 850 hPa (U850, V850).

The manuscript is structured as follows: Section 2 describes the datasets, study region, and predictor selection; Section 3 outlines the methodology; Section 4 presents the results and discussion; and Section 5 provides the conclusions.

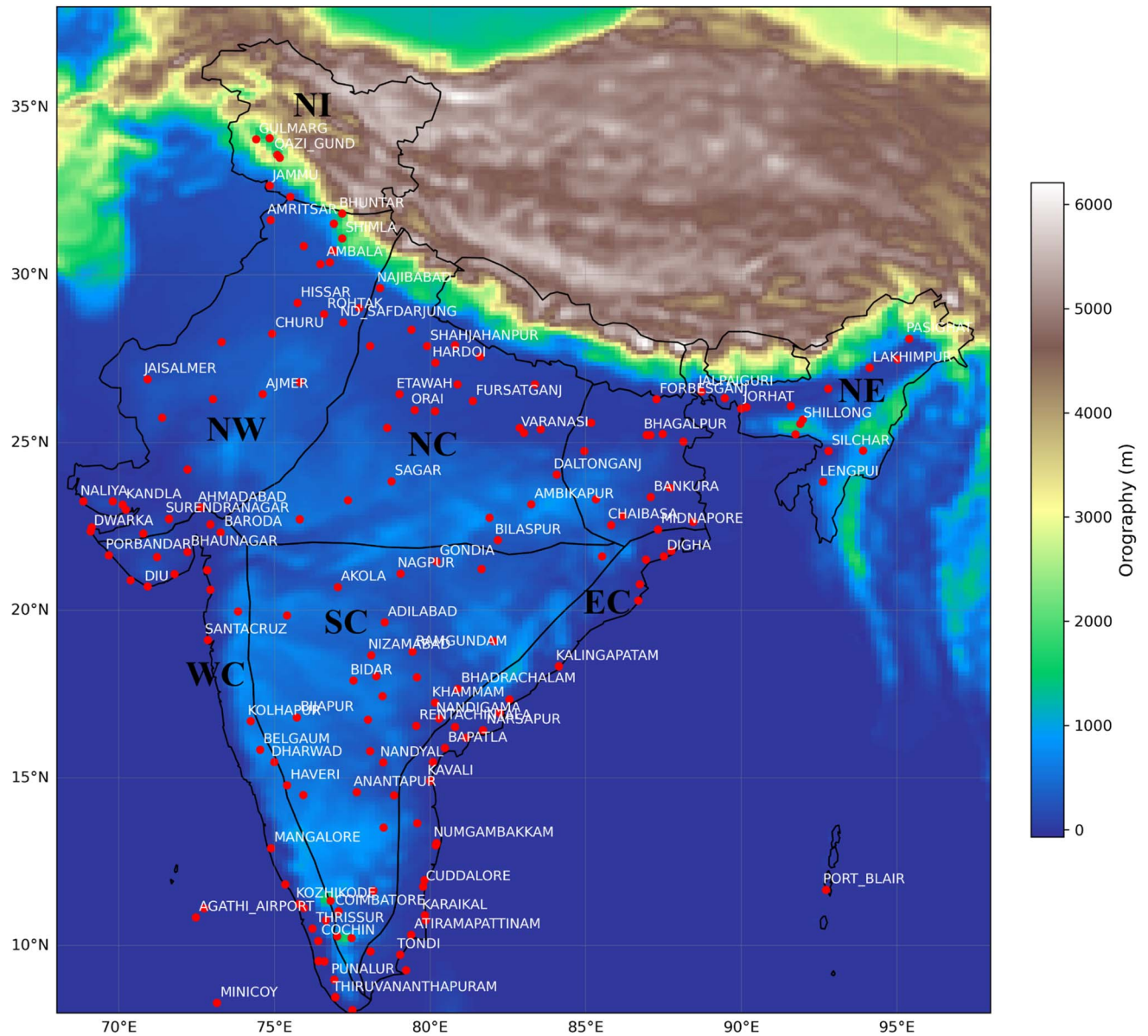

Fig. 1 Orography of the Indian land region showing the seven temperature-homogeneous regions and locations of the 179 IMD stations (red circles).

Table 1 List of the surface variables used as predictors in the training set for bias correcting T_{max} and T_{min} forecasts from NCUMG

Surface variable	Variable description	Unit
APCP_24	24 hour accumulated precipitation	mm
APCP_12z	Accumulated precipitation valid at 12z	mm
DSWRF	Downward shortwave radiation at the surface	W m^{-2}
PRMSL_msl	Mean sea level pressure reduced to MSL	Pa
RH2m	Relative humidity at 2m	%
T2m	Temperature at 2 m	°C
TSOIL1m	Soil temperature 0–0.1 m below the ground	K
U10m	Zonal wind at 10 m	m s^{-1}
V10m	Meridional wind at 10 m	m s^{-1}
ULWRF_surf	Upward longwave radiation flux at the surface	W m^{-2}
TMIN_OBS	Observed daily minimum temperature	°C
TMAX_OBS	Observed daily maximum temperature	°C
AVG_WIND_OBS	Observed average daily wind speed	m s^{-1}
RF_OBS	Observed daily rainfall	mm
TMAX_FCST	Raw model forecast of daily maximum temperature	°C

2. Study area and datasets

This section provides a detailed description of the observed and forecast datasets used in this study, the study regions, and the selection of predictor variables.

2.1 Study area

There is a pronounced spatial variability in temperature over the Indian land region due to its complex geography and climatic diversity. To account for this heterogeneity, IMD has divided the country into seven temperature-homogeneous regions based on long-term climatology.[38–40] These regions are as follows: North (NI), North-Central (NC), North-East (NE), North-West (NW), East Coast (EC), West Coast (WC), and South-Central (SC) (Fig. 1). The NI mainly covers India's northernmost states (Jammu, Kashmir, and Leh). In this region, temperatures range from very cold in winter to mild during summer. The NC region represents central India and is characterized by hot summers and cold winters with frequent heatwaves. The NE region covers the eastern and northeastern parts of the country and experiences mild to cold winters and warm summers influenced by the complex terrain. The NW region experiences very hot summers (42–45 °C) and mild winters (14–18 °C), while the EC and WC regions, close to the coasts, experience warm conditions year-round (summer 28–34 °C and winter 18–22 °C). The SC region covers the peninsular interior and experiences moderate to high summer temperatures (T_{max} ~34–38 °C) with relatively mild winters (T_{min} ~18–22 °C). Fig. 1 also displays the geographical locations of the 179 IMD stations used in this study, along with the terrain. The number of stations is almost uniform across all seven regions, except for NI, which has only 6.

2.2 Observations

Daily station-level T_{max} and T_{min} data were obtained from the IMD archives for 2019–2024. Although IMD operates a large observational network,[41] only 179 stations possessed continuous daily records of T_{max} and T_{min} spanning this entire period, which coincides with the availability of NCUM forecasts. Standard quality control procedures were applied to the observed dataset following Kothawale *et al.*[42] and Srivastava *et al.*[43] A compound outlier criterion was adopted: a station day was flagged for removal only if the observed temperature simultaneously fell below the 1st or above the 99th percentile of the station climatological distribution, and the daily anomaly relative to adjacent days exceeded ±10 °C. Values satisfying the percentile criterion alone but not the anomaly criterion were retained, ensuring that physically plausible extreme temperatures associated with heatwave and cold wave events are preserved in the dataset. Cases where T_{max} was lower than T_{min} were treated as missing and filled using the approach of Kothawale *et al.*[42] In this method the replacements are based on the mean of neighboring days when the surrounding period (3–4 days) exhibited relatively uniform conditions. After quality control, daily T_{max} and T_{min} records for 179 stations were retained for the period 2019 to 2024.

In Fig. 2, panels (a) and (g) depict the mean observed daily T_{max} and T_{min}, respectively, derived from the IMD high resolution daily gridded temperature dataset at 0.5° × 0.5° resolution (~50 km) developed by Srivastava *et al.* (2009).[43] Although this dataset is available from 1969 onwards, the present study utilizes only the 2019–2024 period, consistent with the training and test data used for ML bias correction. For T_{max} (Fig. 2a), the highest climatological values are observed over the NW, NC, and SC regions, showing the high temperatures during the pre-monsoon season. On the other hand, comparatively lower temperatures are observed over NI, EC, and WC due to their elevations and maritime influences. The T_{min} climatology (Fig. 2g) exhibits a strong north–south gradient, with lower temperatures over the NI, NW, NE, and NC regions. The SC, EC, and WC regions show higher T_{min} values because they are closer to the coast.

2.3 NCUM forecast data

At NCMRWF, the NCUMG (V7) model is the operational weather forecasting system with a horizontal resolution of ~12 km and

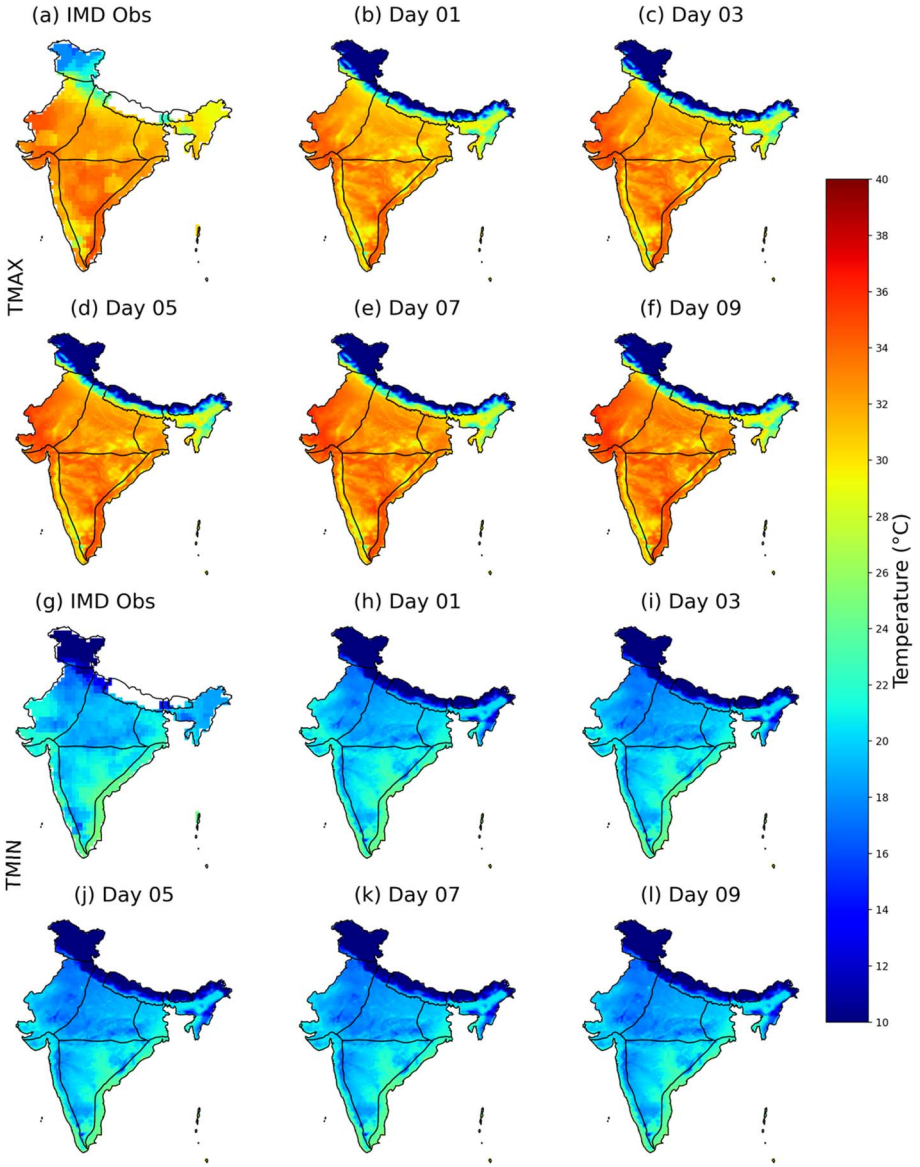

Fig. 2 IMD observed (a and g) and NCUM forecast T_{max} (b–f) and T_{min} (h–l) climatology during the period 2019–2024.

70 vertical levels, extending to 80 km. It uses the "ENDGame" dynamical core,[44] which provides improved accuracy of the solution of primitive model equations and reduced damping. The Hybrid 4-Dimensional Variational (4D-Var) data assimilation scheme is used to create the analysis of the NCUMG model.[45] A detailed description of the NCUM Hybrid 4D-Var system is given in Kumar *et al.*[46–48]

2.3.1 Nature of biases in NCUMG. Fig. 2a and (g) show the mean observed daily T_{max} and T_{min}, respectively, while Fig. 2b–f and (h–l) illustrate the corresponding NCUMG climatological forecasts for lead times of Day-1, Day-3, Day-5, Day-7, and Day-9, for the period 2019–2024. Overall, NCUMG reproduces the large-scale spatial patterns of the observed temperature climatology reasonably well across all lead times. The pronounced north–south contrast, with higher temperatures over the NC, NW, and SC regions and relatively lower temperatures over the NE, NI, and coastal regions, is well captured by the model. However, systematic biases are seen in both T_{max} and T_{min}, with their magnitude and spatial structure evolving with forecast lead time.

For T_{max}, the Day-1 forecast (Fig. 2b) already shows an underprediction of approximately 2–3 °C over the NW and NC regions, which are prone to heatwaves during the pre-monsoon season. This cold bias persists and increases with lead time, reaching approximately 4–6 °C over NC and SC by Day-9 (Fig. 2c–f). Over NI, a cold bias is observed across all lead times. The EC and WC regions show comparatively smaller T_{max} biases, where the moderating influence of the ocean limits temperature variability and constrains forecast errors relative to the drier continental interior.

For T_{min}, the Day-1 forecast (Fig. 2h) shows a warm bias of approximately 1–2 °C over the NC region, which persists and

intensifies at longer lead times, reaching approximately 2–4 °C over inland regions by Day-9 (Fig. 2i–l). The NW region shows a growing cold bias in T_{min} with lead time. As for T_{max}, EC and WC regions again show comparatively smaller T_{min} biases, due to the moderating oceanic influence on nocturnal temperatures in these coastal zones.

While the annual-mean bias patterns provide useful information about errors in T_{max} and T_{min} forecasts from NCUMG, they do not convey any information about the seasonal behavior of temperature. Temperature forecast errors exhibit seasonal variations, particularly over India, where surface processes, land–atmosphere coupling, and boundary-layer dynamics differ substantially between summer and winter. To better understand these effects, we have also examined seasonal mean biases for T_{max} during the pre-monsoon season (MAMJ) in Fig. 3a–e and for T_{min} during the winter (DJF), as shown in Fig. 3f–j. The figure shows that during MAMJ, T_{max} forecasts exhibit a warm bias over large parts of the NW, NC, and SC regions, with the bias increasing from Day-1 (∼0.5–1 °C) to Day-9 (∼4–5 °C). This contrasts with the annual-mean cold bias in T_{max} (Fig. 2). Conversely, during DJF, T_{min} forecasts show a widespread cold bias, particularly over NI, NC, and NW regions, which intensifies with lead time. As before, this behaviour differs from the annual-mean warm bias in T_{min}. The bias in T_{min} also increases (*i.e.*, becomes more pronounced) with increasing lead time from Day-1 (∼−1 °C) to Day-9 (∼−4 to −5 °C). A notable exception is the NI region, which exhibits a strong negative bias of approximately 3–5 °C in both MAMJ T_{max} and DJF T_{min} across all lead times. This cold bias over NI is particularly persistent and does not show the same lead time dependence seen in other regions, suggesting a systematic underestimation likely linked to the model's representation of orographic and boundary layer processes in this complex terrain region.

Overall, these results demonstrate that annual-mean T_{max} and T_{min} biases result from opposing seasonal error characteristics. These temperature biases are closely linked to errors in the model's representation of the underlying thermodynamic processes, including surface energy balance, boundary-layer evolution, soil–atmosphere interactions, and nocturnal radiative cooling, which vary strongly between seasons over India.[21,49–51]

2.4 Predictor selection

Predictor selection plays a key role in developing robust ML models by capturing relevant physical relationships while avoiding redundancy. An overly large predictor set can lead to overfitting[52,53] and multicollinearity,[54,55] thereby reducing computational efficiency. In this study, predictors from the NCUMG forecast were selected based on their ability to minimize RMSE while preserving essential inter-variable relationships.

The predictor set (Fig. 4) chosen in this study is presented in Tables 1 and 2, this includes accumulated precipitation over 24 hours and at 12 UTC (APCP_24, APCP_12z), surface and top-of-atmosphere shortwave and longwave radiation (DSWRF, ULWRF_surf, ULWRF_top), and large-scale circulation indicators such as sea-level pressure (PRMSL_msl) and 500 hPa

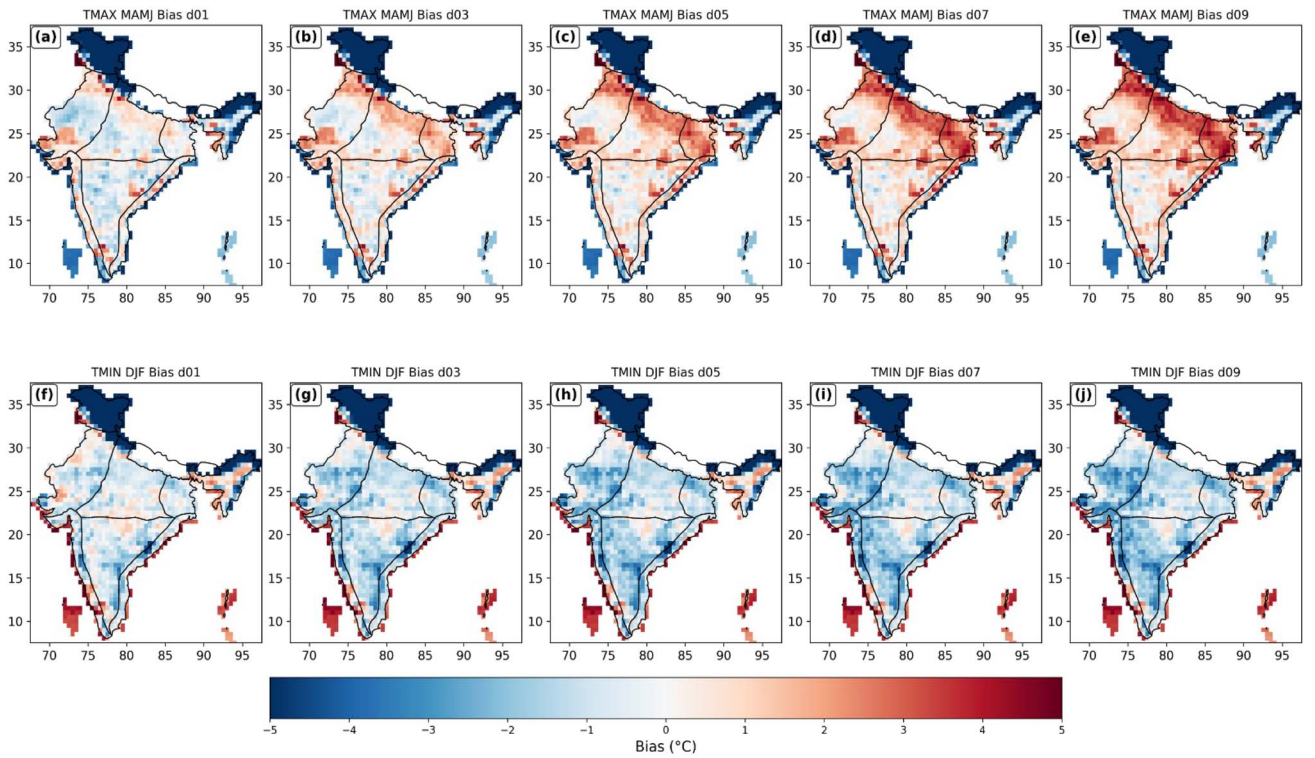

Fig. 3 Mean biases (2019–2024) in the NCUMG forecasts March–June (a–e) for T_{max} and December–February for T_{min} (f–j).

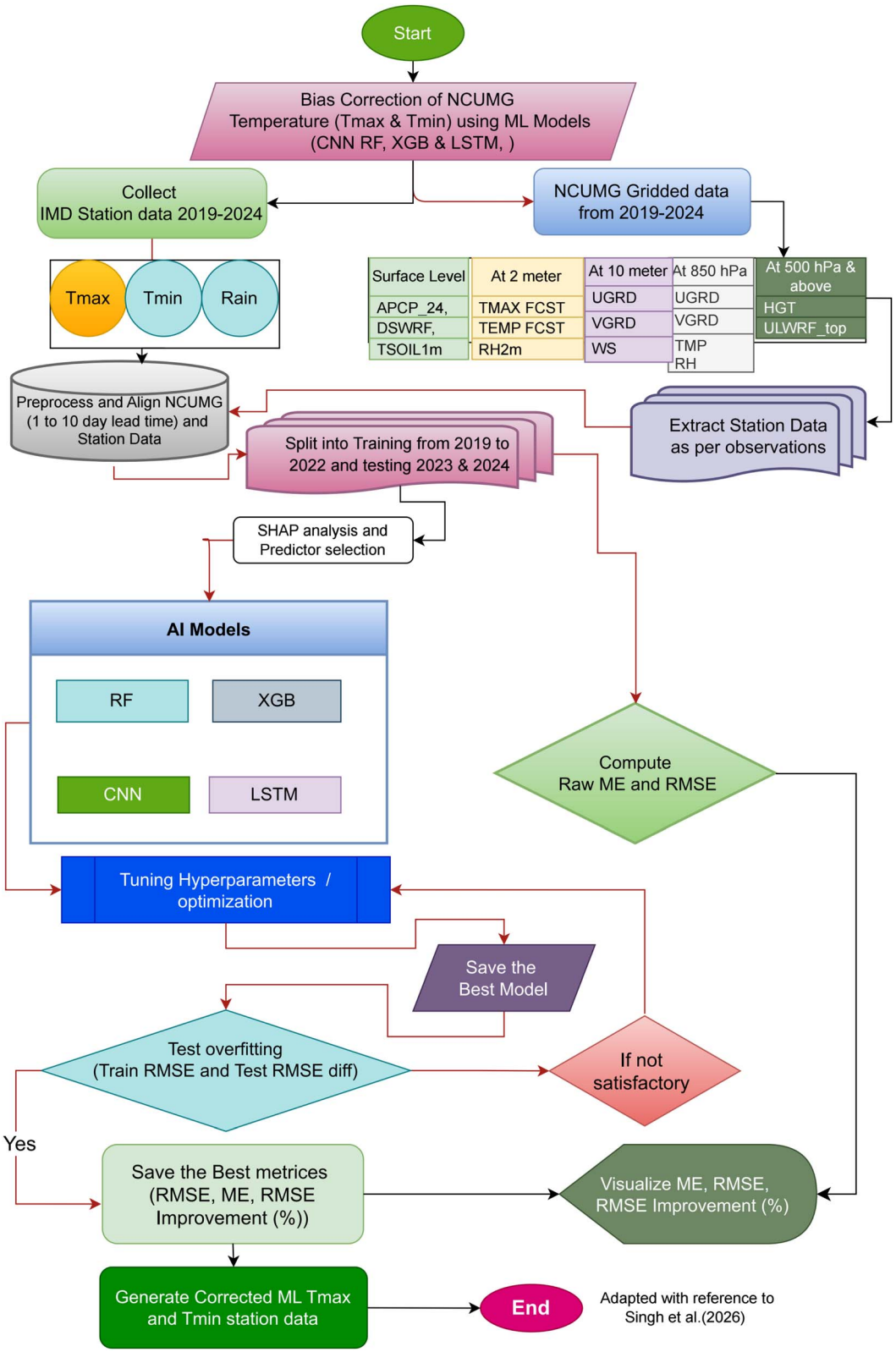

Fig. 4 Flowchart of the methodology.

geopotential height (HGT500). Parameters such as relative humidity at 2 m and 850 hPa (RH2m, RH850), air temperature at 2 m and 850 hPa (T2m, T850), soil temperature below the ground 0–0.1 m (TSOIL1m), and horizontal wind components at 10 m and 850 hPa (U10m, V10m, U850, V850) collectively describe boundary-layer moisture, stability, and advection

Table 2 List of the upper air variables used as predictors in the training set for bias correcting T_{max} and T_{min} forecasts from NCUMG

Upper air variable	Level	Variable description	Unit
HGT500	500 hPa	Geopotential height	m
RH850	850 hPa	Relative humidity	%
T850	850 hPa	Air temperature	K
U850	850 hPa	Zonal wind	m s^{-1}
V850	850 hPa	Meridional wind	m s^{-1}
ULWRF_top	TOA	Upward long wave radiation flux at the top of the atmosphere	W m^{-2}

patterns. Additional station-based predictors, such as daily T_{min} (TMIN_OBS), T_{max} (TMAX_OBS), average wind speed (AVG_WIND_OBS), and observed rainfall (RF_OBS), are also used to bias-correct the target variables, *i.e.*, T_{max} and T_{min} from NCUMG forecasts. While correcting the T_{max} forecasts, TMIN_OBS is used as one of the predictors and *vice versa*.

The NCUMG forecast dataset, along with the observations, is then split into training and test sets, with the training set including forecast and observation data from 2019 to 2022 for each station. T_{max} and T_{min} forecasts from the remaining years 2023 and 2024 are used as the test set to evaluate model performance after bias correction.

3. ML techniques used for bias correction

The methodology adopted in this study is illustrated in Fig. 4. Four ML methods were applied for bias correction: RF, XGB, CNN, and LSTM. All ML models were trained independently for each station and lead time combination, resulting in a separate trained model for each of the 179 stations across lead times Day-1 to Day-10. This station and lead time specific training strategy ensures that each model captures the localised bias characteristics of individual stations and the lead time dependent error growth of the NCUMG. The selection of these methods is motivated by Singh *et al.*,[21] who evaluated a broader set of models including RF, XGB, CNN, Support Vector Machines (SVM), and Multiple Linear Regression (MLR) for correcting station-level T_{max} and T_{min} biases in the IMDAA reanalysis over India. It was found in their study that RF, XGB, and CNN consistently outperformed MLR and SVM in handling the nonlinear and multivariate nature of temperature biases. Since both IMDAA and NCUMG are based on similar underlying model physics and exhibit comparable error characteristics, RF, XGB, and CNN are retained in the present study to assess their effectiveness for bias correction of NCUMG temperature forecasts. Python was used to implement the above-mentioned ML methods, with NumPy and Pandas for data handling, scikit-learn[56] for traditional ML algorithms, and TensorFlow/Keras for deep learning model construction, training, and callbacks.

RF and XGB are well established ensemble and boosting methods widely used for NWP bias correction.[33,56–58] For RF, hyperparameters were selected *via* grid search over the number of trees (n_estimators ∈ {100, 200, 300}), maximum depth ∈ {10, 15, 20}, minimum samples per leaf ∈ {5, 10, 15}, and maximum features ∈ {2, 3, 4}, selecting the combination that minimised validation RMSE for each station and lead time. For XGB, the learning rate, maximum depth, and related tree parameters were tuned using 5-fold cross-validation with early stopping over up to 80 boosting rounds, after which the final model was retrained on the full training set using the optimal parameter combination, with L1 and L2 regularisation applied throughout to prevent overfitting.

The CNN architecture consists of two one-dimensional convolutional layers with 32 and 16 filters respectively (kernel size 3), followed by a dense layer with 32 units and a linear output neuron. A dropout rate of 0.2 was applied after the first convolutional block and the dense layer, with early stopping and learning rate reduction on validation loss used to control overfitting. The CNN was trained using the Adam optimiser with a learning rate of 1×10^{-3}, a batch size of 32, and up to 100 epochs.[21,59] The LSTM architecture comprises one LSTM layer with 64 hidden units, followed by a dense output layer (32 units and a final linear neuron), with a dropout rate of 0.2 applied between the LSTM and dense layers; hyperparameters (number of units, dropout, learning rate 0.001, batch size 32 and up to 100 epochs with early stopping) were selected by minimising cross-validated RMSE on the training set.[36,37] For all models, hyperparameters were optimised on the 2019–2022 training period and final performance evaluated on the independent 2023–2024 test set.

3.1 Verification metrics

The performance of the raw NCUM forecast dataset and ML models was compared using well-known metrics, such as ME (over- and underestimation) and percentage improvement in RMSE (magnitude of error). In addition to these metrics, we have also analyzed the performance of the bias correction models in predicting extreme T_{max} and T_{min}, using categorical verification metrics like:

Equitable Threat Score (ETS) evaluates model performance relative to random chance, with values ranging from −1 to 1 (perfect skill).[7,60]

Heidke Skill Score (HSS)[61] measures overall forecast skill, ranging from $-\infty$ to 1 (perfect skill). For deterministic forecasts, HSS is mathematically equivalent to Cohen's Kappa (CK).

Relative Economic Value (REV): the value of forecasts depends on their ability to improve decisions, not just accuracy.[62] Bias correction reduces errors but may not always increase usefulness.[63] The REV[62] quantifies usefulness across cost–loss ratios: REV ranges from −1 (worse than using no forecast) to 1 (perfect forecast).

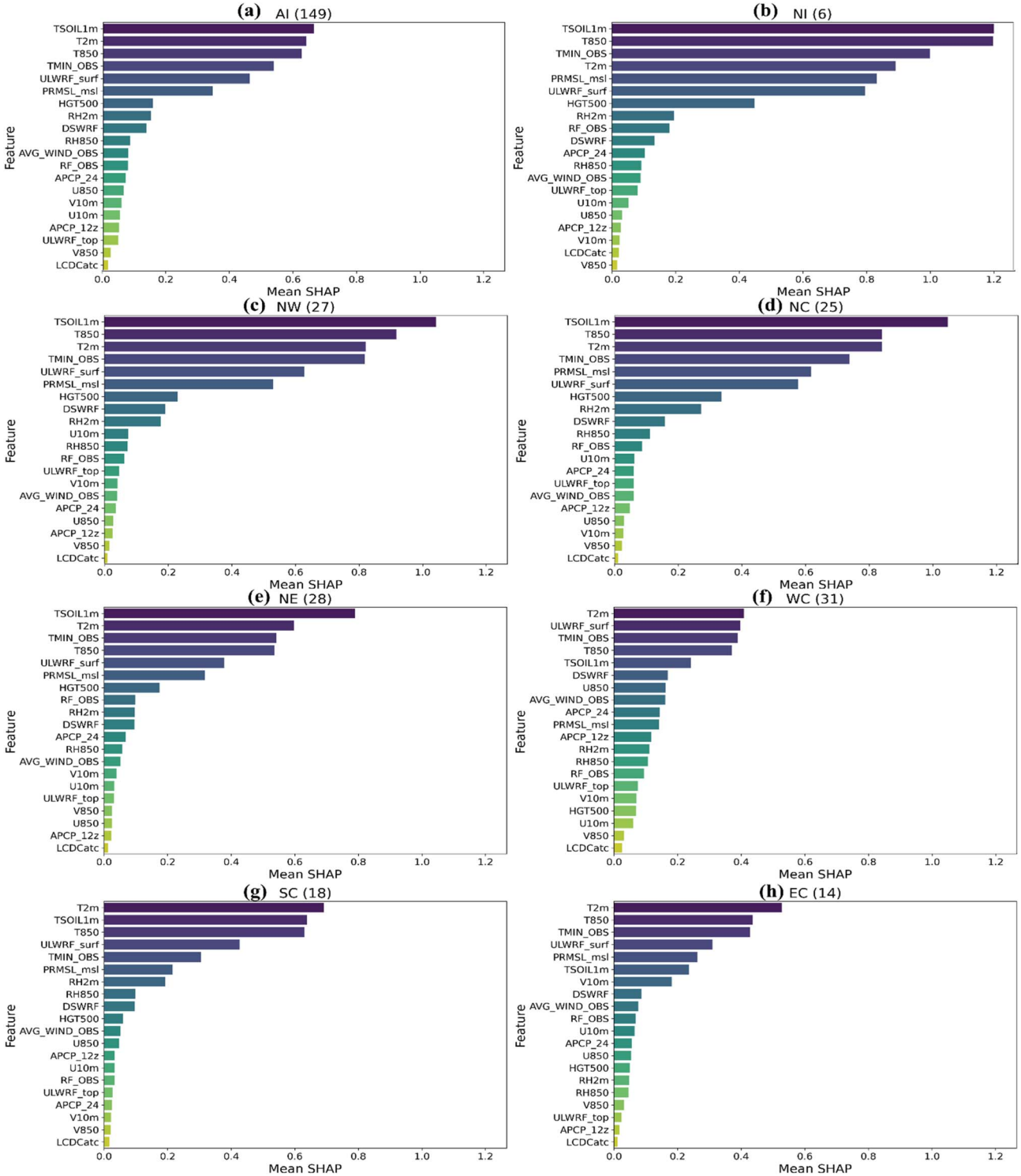

Fig. 5 (i) Region-wise mean SHAP (Shapley Additive Explanations) featuring the importance of predictor variables used for bias correction of NCUMG daily TMAX forecasts during 2023–2024 across All India (a) and the seven homogeneous regions (b–h); and (ii) region-wise mean SHAP (Shapley Additive Explanations) featuring the importance of predictor variables used for bias correction of NCUMG daily TMIN forecasts during 2023–2024 across All India (a) and the seven homogeneous regions (b–h).

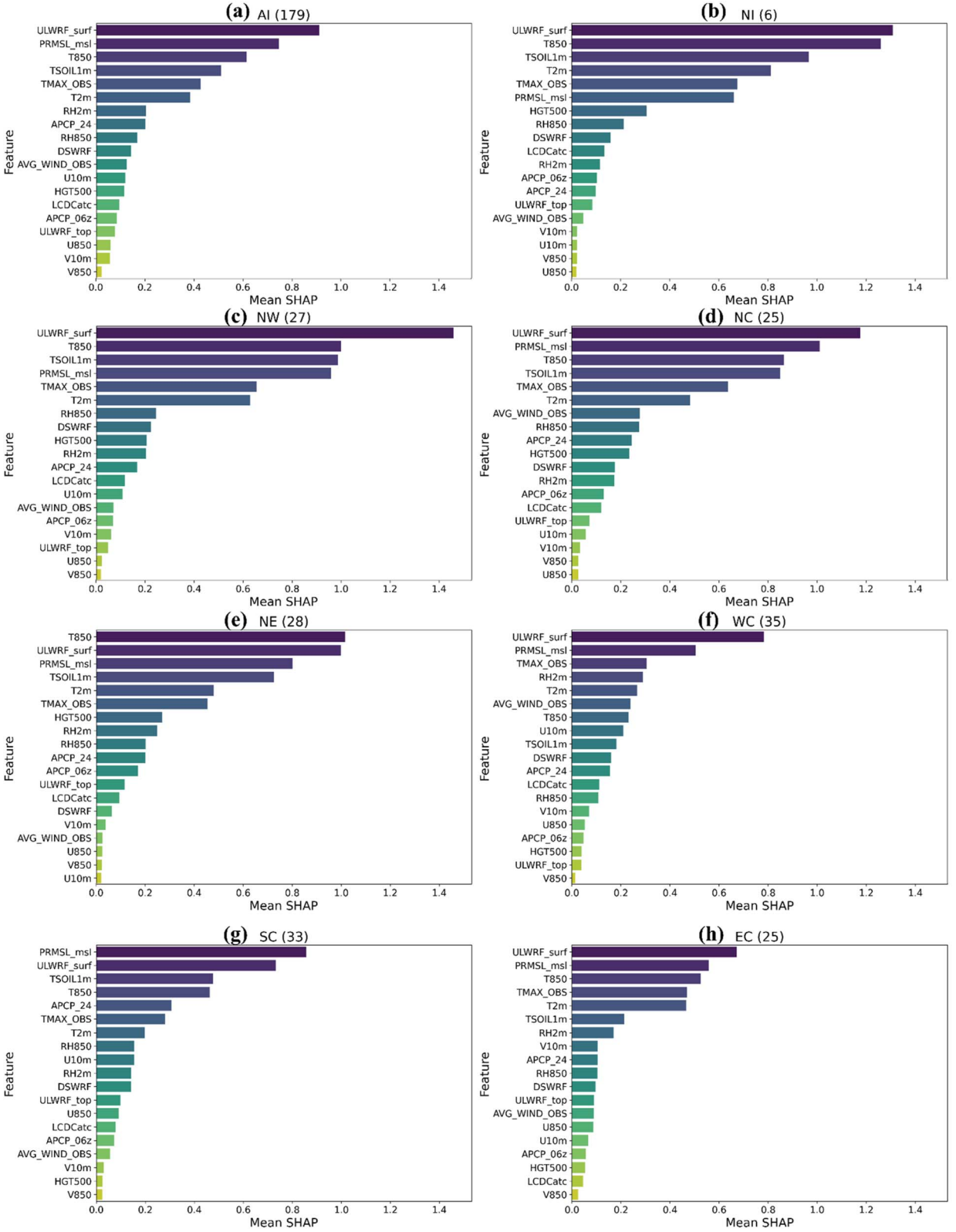

Fig. 5 (*contd.*)

3.2 Interpretation of machine-learning bias correction using SHAP

In this study, we also used SHAP (Shapley Additive Explanations) to quantify each predictor's contribution to bias correction at individual stations. This helps explain why the ML model applies a particular correction by linking the adjustment to specific meteorological variables, thereby providing a physically meaningful understanding of the bias patterns.[64–66]

3.2.1 SHAP-based drivers of T_{max} bias correction. The SHAP analysis reveals large-scale similarities but some regional differences in the predictors controlling T_{max} bias correction across India (Fig. 5). Across all regions, T2m and T850 consistently rank among the most influential predictors, confirming that boundary layer thermodynamics and large-scale thermal structure play fundamental roles in regulating daytime T_{max} regardless of regional climate. The persistent importance of observed T_{min} further indicates that nocturnal thermal conditions influence subsequent daytime heating through soil heat retention and boundary layer recovery.

In inland continental regions, NW, NC, and SC, TSOIL1m emerges as a dominant predictor along with T850, indicating strong land–atmosphere coupling under dry conditions where daytime T_{max} is closely tied to prior soil thermal state. These regions also show appreciable contributions from radiative fluxes, consistent with clearer skies and stronger surface energy exchanges. In contrast, coastal regions (WC and EC) show reduced TSOIL1m importance and greater reliance on T2m, ULWRF_surf, and observed T_{min}, reflecting the moderating influence of maritime air masses and cloud-modulated radiative processes. The mountainous NI region shows both TSOIL1m and T850 as jointly dominant, indicating orographic boundary layer trapping and free tropospheric temperature advection in complex terrain. The NE region shows relatively higher contributions from radiative and humidity-related variables, consistent with frequent cloud cover and humid boundary layer conditions.

Regions with a higher proportion of urban stations such as NC and NW show greater importance of observed T_{max} and T_{min}, showing thermal persistence associated with urban heat island effects where built surface heat retention makes prior day temperatures a strong predictor of subsequent maxima. Rural and forested regions such as NE show comparatively lower observed temperature importance, consistent with more efficient overnight thermal recovery under natural land cover.

3.2.2 SHAP-based drivers of T_{min} bias correction. The SHAP analysis for T_{min} is presented in Fig. 5b. Across all regions, ULWRF_surf and T850 are consistently the dominant predictors, confirming that nocturnal radiative cooling and large-scale atmospheric thermal structure are the primary controls for T_{min}. The frequent importance of PRMSL_msl further reflects the influence of synoptic-scale circulation on nighttime stability, *i.e.*, anticyclonic conditions favour clear skies and enhanced radiative cooling, while low-pressure systems promote cloudiness and moisture that suppress nocturnal cooling. Misrepresentation of these pressure patterns in the model may therefore directly translate into T_{min}.

In inland regions, NW, NC, and SC, T_{min} bias correction is strongly influenced by ULWRF_surf, PRMSL_msl, T850, and TSOIL1m, showing the combined role of strong radiative cooling under clear skies and soil heat storage and release during nighttime hours in dry climates. Urban stations in NC and NW additionally show higher importance of observed T_{max}, consistent with urban heat island effects where daytime heat retention in built surfaces elevates nocturnal temperatures. In NI, ULWRF_surf and T850 dominate, due to strong nocturnal radiative cooling and free tropospheric temperature advection in complex mountainous terrain where boundary layer stability is strongly modulated by orography. Across NE, humidity and precipitation-related predictors contribute along with ULWRF_surf and PRMSL_msl, because of frequent cloud cover and humid boundary layer conditions that can modulate nocturnal cooling.

Along both WC and EC, T_{min} corrections are primarily governed by ULWRF_surf and PRMSL_msl with additional contributions from observed T_{max} and T2m, indicating strong coupling between daytime heating and nighttime temperatures under maritime conditions where enhanced moisture suppresses radiative cooling and reduces the role of soil-related variables.

4. Results and discussion

This section presents detailed verification of the bias-corrected NCUMG forecasts obtained from each of the methods defined above. The verification uses standard metrics, such as RMSE and ME, to assess the magnitude of errors before and after bias correction in the test set. We have also conducted a categorical verification of the test set using verification scores such as ETS. Finally, to assess improvement in the bias-corrected results, we have also calculated the Relative Economic Value (REV) for both the T_{max} and T_{min} forecasts. While all lead times were analysed, some of the results are presented for the All-India region at representative lead times (Day 1, 3, 5, 7, and 9) to maintain clarity and brevity. The detailed results are presented below.

4.1 Improvement in ME, RMSE, and correlation

To evaluate the performance of the bias-correction methods, ME (over- or underestimation) and RMSE, which measures the magnitude of error, were analysed for the raw and post-processed forecasts. Fig. 6 and 7 show boxplots of ME for T_{max} and T_{min}, respectively, for Day 1–10 forecasts over All India (AI) and the seven homogeneous regions.

4.1.1 ME in T_{max}. For T_{max} (Fig. 6), the raw NCUMG forecasts exhibit systematic biases that generally increase with lead time, with both the magnitude and sign varying across regions. Over AI, NW, NC, and SC, the raw forecasts show a gradual warm bias at longer lead times, whereas persistent cold biases are seen over the EC and WC regions.

All bias-correction methods reduce the ME across regions and lead times, although their effectiveness varies. For AI (Fig. 6a), LSTM yields the median ME closest to zero across all lead times (typically within ±0.1 °C). In contrast, RF and XGB

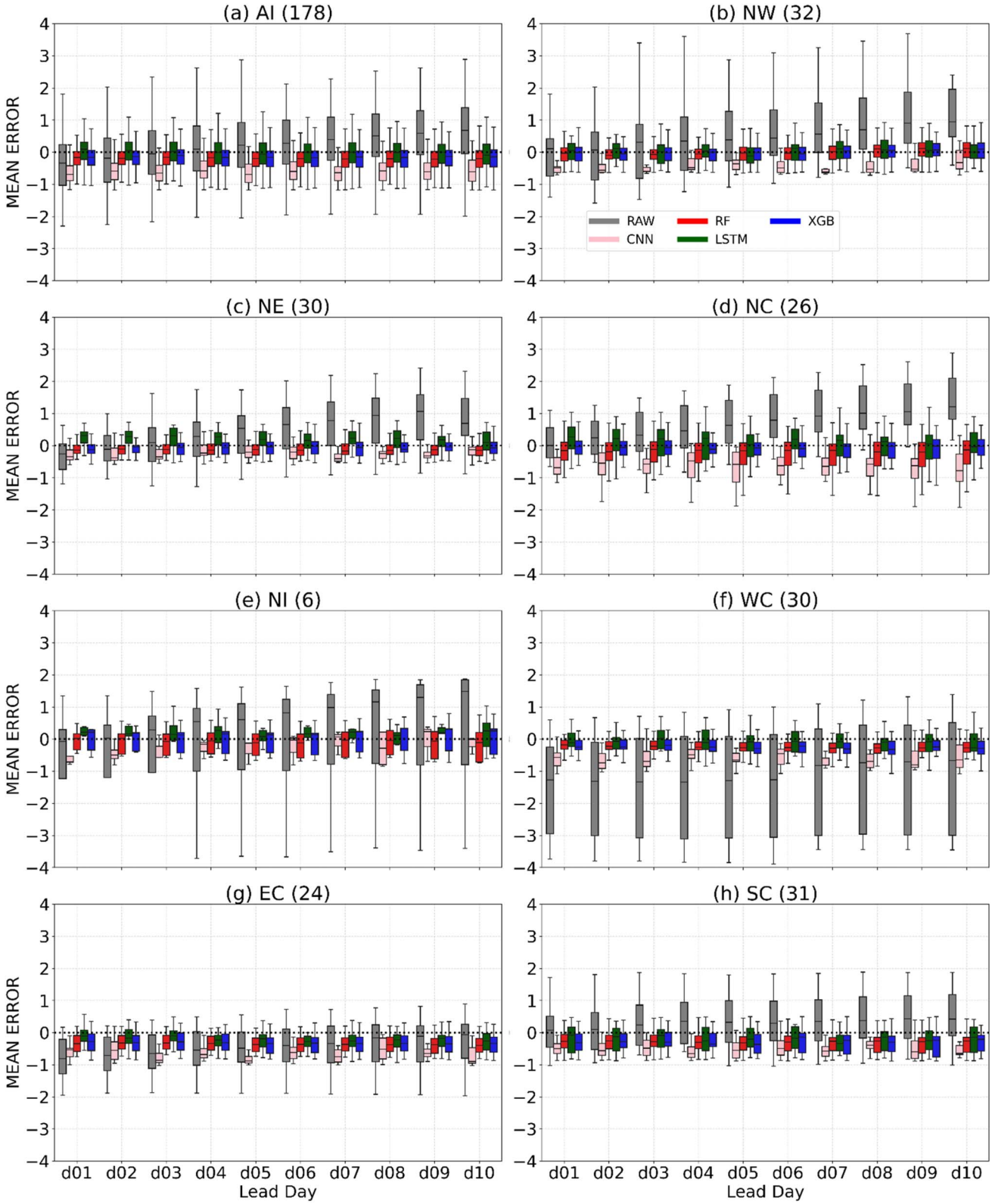

Fig. 6 Regionwise mean error (°C) (a) all India region (AI), (b) northwest (NW), (c) north east (NE), (d) north central (NC), (e) north India (NI), (f) west coast (WC), (g) east coast (EC), and (h) south central (SC) for NCUM model TMAX forecast during 2023–2024.

show a small but consistent negative bias, with median ME generally around −0.2 to −0.4 °C, but with a tighter interquartile range (IQR) of about ±0.3–0.5 °C compared to the raw model. CNN consistently underestimates T_{max} with median ME often between −0.5 and −0.8 °C and exhibits a wider spread, with the IQR frequently exceeding ±0.8–1.0 °C, indicating larger

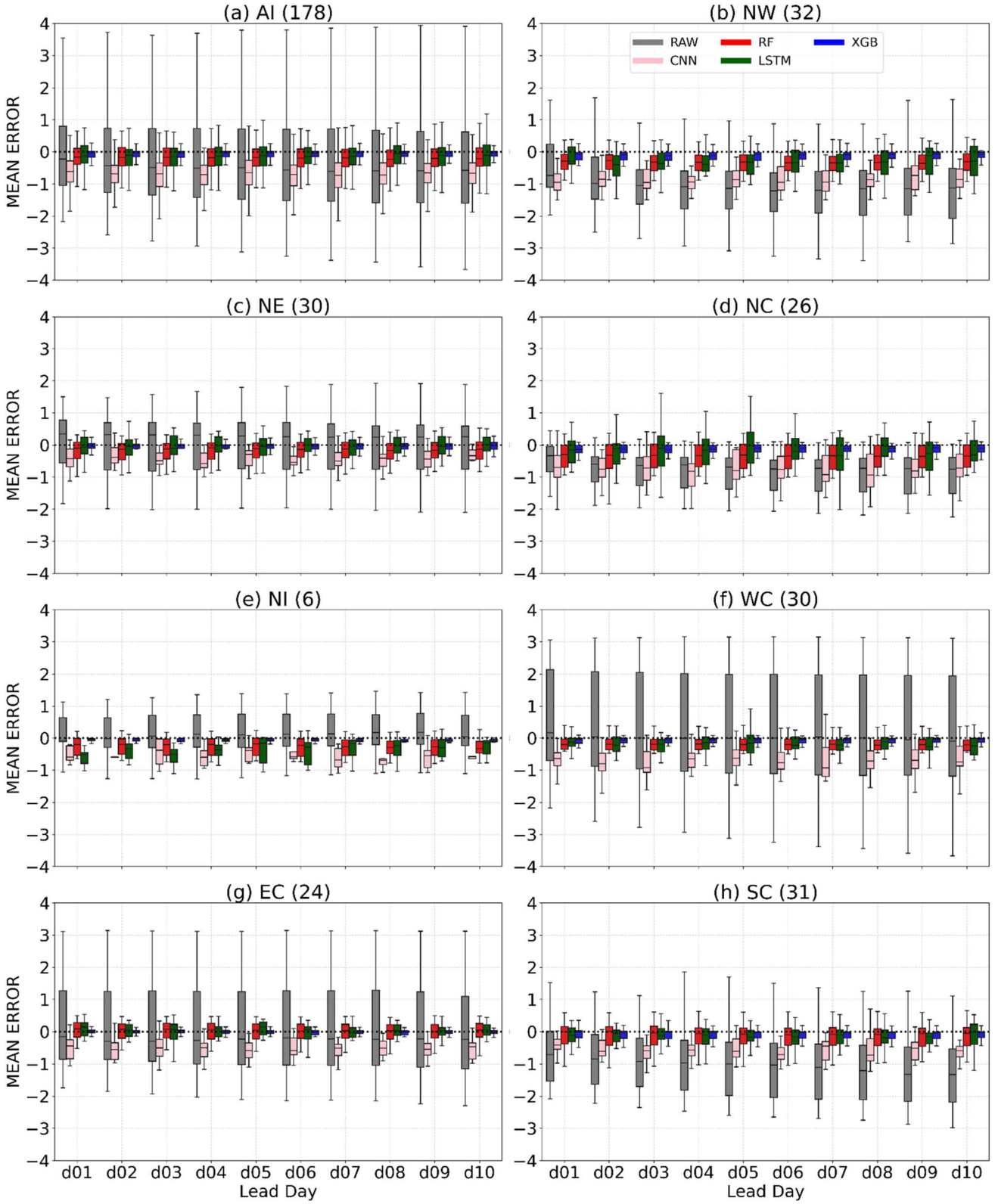

Fig. 7 Regionwise mean error (°C) (a) all India region (AI), (b) northwest (NW), (c) north east (NE), (d) north central (NC), (e) north India (NI), (f) west coast (WC), (g) east coast (EC), and (h) south central (SC) for NCUM model TMIN forecast during 2023–2024.

variability in the residual errors. Over NW (Fig. 6b), RF, XGB, and LSTM show comparable median ME values close to zero (within ±0.2 °C). However, XGB exhibits the smallest spread, with an IQR of about ±0.3–0.4 °C, compared to RF and LSTM where the spread is slightly larger (±0.5–0.7 °C), indicating more consistent bias correction by XGB across stations and lead

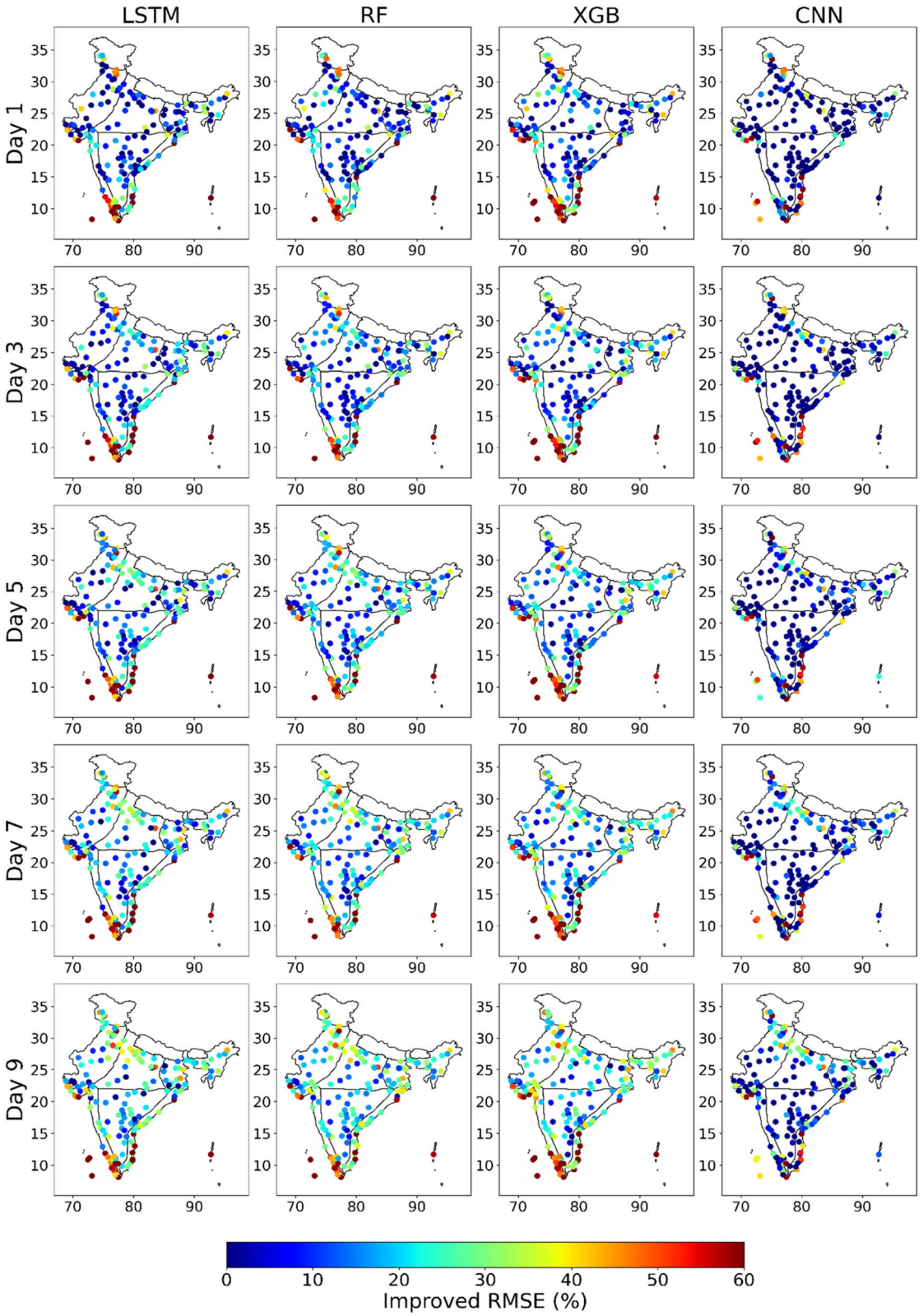

Fig. 8 Improvement in RMSE (%) for NCUM model TMAX forecast during 2023–2024.

times. In the NE region (Fig. 6c), XGB performs best, aligning the median ME close to zero, while RF underestimates and LSTM overestimates T_{max}. A similar behaviour is seen over NC and NI (Fig. 6d and e), where XGB most effectively centres the ME around zero, whereas RF shows underestimation and LSTM overestimation with comparatively larger spread. The raw

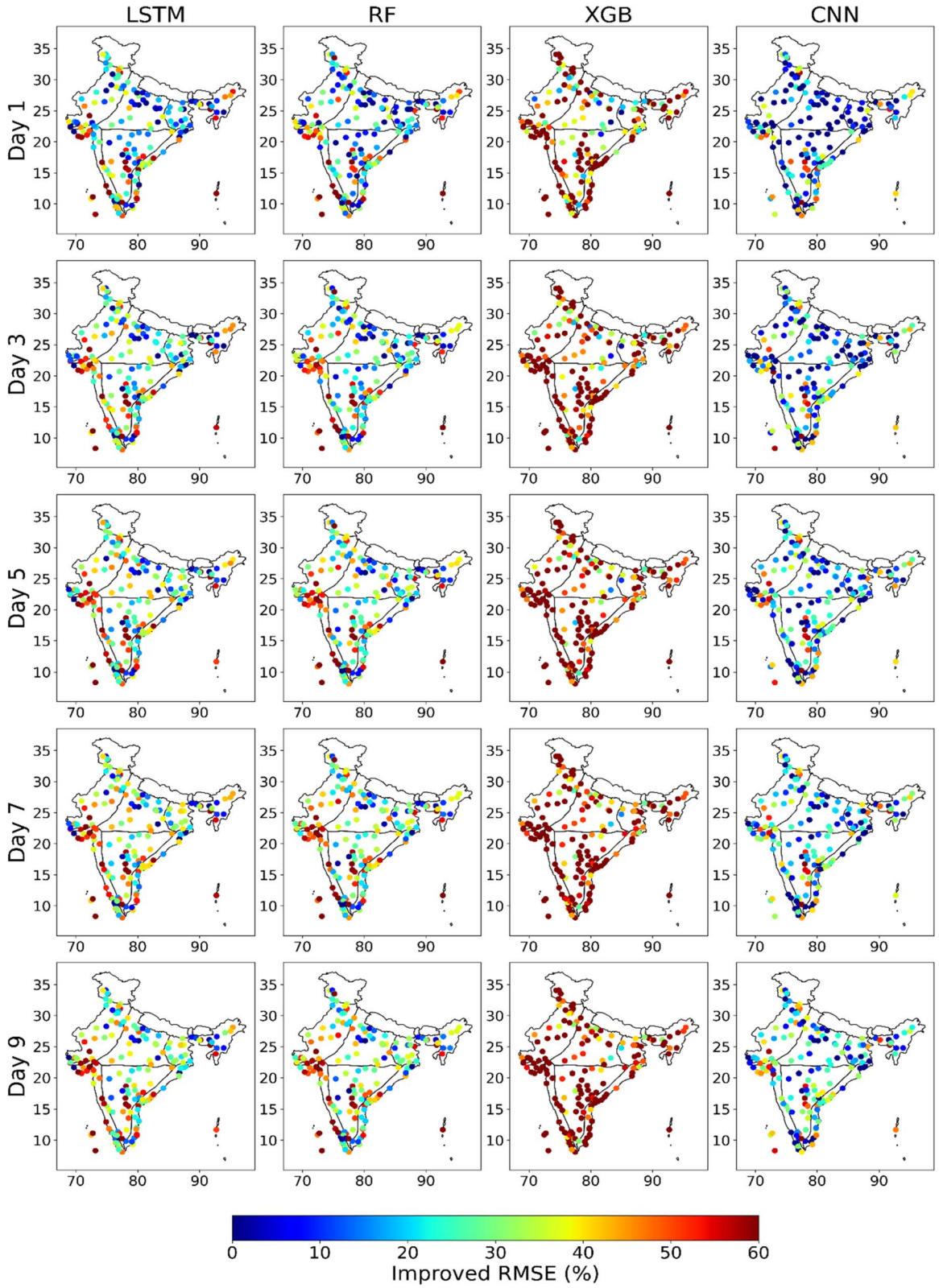

Fig. 9 Improvement in RMSE (%) for NCUM model TMIN forecast during 2023–2024.

NCUMG forecasts exhibit a clear warm bias, with the median ME increasing from about +0.2 °C at Day-1 to nearly +1.0 °C by Day-10. After bias correction, all ML methods shift the median ME below zero, indicating over-correction into a cold bias. Among them, LSTM remains closest to zero (≈ -0.1 to -0.3 °C), while RF and XGB show slightly stronger negative biases

(≈−0.2 to −0.5 °C). CNN exhibits the largest cold bias and wider spread (<−0.5 °C). Although the ML methods substantially reduce the spread compared to the raw forecasts, none is able to fully remove the bias over this region due to this tendency to overshoot into negative ME. In the coastal regions, WC and EC, the raw forecasts show strong cold biases. Over WC, LSTM performs best by bringing the median ME close to zero, while the other ML methods continue to underestimate T_{max}. Across EC, all ML methods reduce bias but continue to show underestimation; LSTM remains closest to zero, particularly at shorter lead times, whereas RF and XGB are farther from zero.

Overall, the ME analysis indicates that ML-based post-processing substantially reduces systematic T_{max} biases, with clear regional differences in model performance. XGB performs best in the inland and northern regions (NW, NE, NC, and NI), where it most effectively centres the ME around zero, whereas LSTM provides better bias reduction in the coastal and southern regions (WC, EC, and SC). Among all methods, CNN consistently performs worse, with larger residual biases and greater spread.

4.1.2 ME in T_{min}. For T_{min} (Fig. 7), the raw NCUMG forecasts display strong regional contrasts in bias structure. Over AI (Fig. 7a), the raw forecasts show a persistent cold bias of about −0.5 to −1.0 °C, which increases with lead time. Similar cold biases are evident over NW, NC, EC, and SC, where the median ME typically ranges between −0.5 and −1.5 °C by Day-10. In contrast, NI, WC, and NE show a clear warm bias, with median ME values between +0.5 and +1.5 °C, indicating overestimation of nighttime temperatures in these regions.

All ML-based methods substantially reduce these biases. XGB consistently brings the median ME closest to zero across regions and lead times, typically within ±0.2 °C, and also shows the smallest spread, with an IQR within ±0.3–0.4 °C. RF and LSTM also reduce bias but tend to retain a slight cold bias (−0.2 to −0.6 °C) across most regions, with a wider IQR of approximately ±0.5–0.8 °C, indicating greater variability compared to XGB.

Unlike T_{max}, where model performance varies by region, T_{min} bias correction is clearly dominated by XGB, which most effectively centers the ME around zero irrespective of region. The raw T_{min} biases also show greater heterogeneity than T_{max}: warm biases over NI, WC, and NE and cold biases over NW, NC, EC, and SC, which may be attributed to the different physical controls on nighttime cooling. These regional contrasts emphasize upon the importance of variable-specific and region-aware bias-correction strategies.

4.1.3 T_{max} RMSE improvement. Fig. 8 indicates that XGB and LSTM yield the largest reductions in RMSE, while CNN provides minimal benefit, consistent with earlier error analyses. The figure also shows that RMSE improvements are modest at short lead times and increase with forecast lead time, consistent with the growth of systematic biases observed in Fig. 2 and 3. This is evident from the Day-1 plot over NC and SC (Fig. 8k), where most stations show limited RMSE improvement (∼10–15%) as the raw forecasts already have relatively smaller errors at short lead times. By Day-5, RMSE improvements strengthened across these regions, with many stations exhibiting gains of ∼20–30%, particularly for LSTM and XGB. By Day-9, there is substantial improvement across several stations, frequently

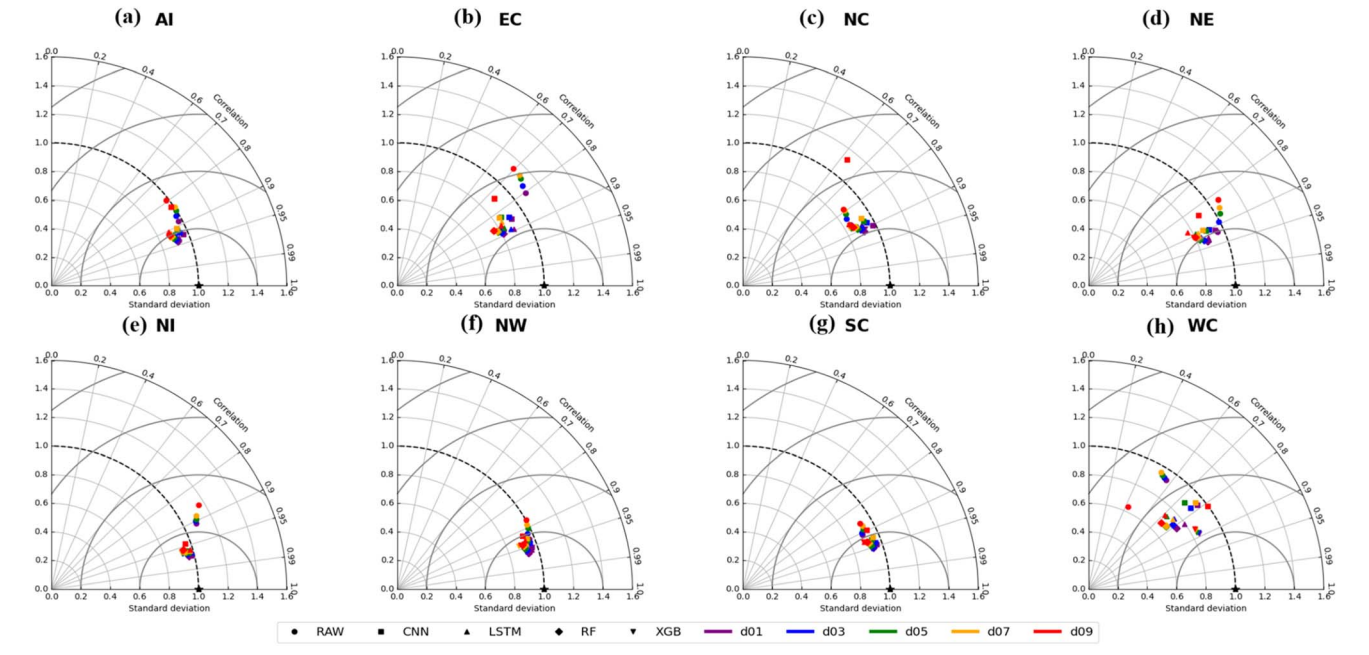

Fig. 10 Taylor diagram analysis of TMAX forecasts showing correlation, normalized standard deviation, and centered root mean square error for the raw NCUMG forecasts and ML-based bias-corrected forecasts (RF, XGB, LSTM, and CNN) at lead times Day-1, Day-3, Day-5, Day-7, and Day-9 over (a) All India (AI), (b) East Coast (EC), (c) North-Central (NC), (d) North-East (NE), (e) North India (NI), (f) North-West (NW), (g) South-Central (SC), and (h) West Coast (WC).

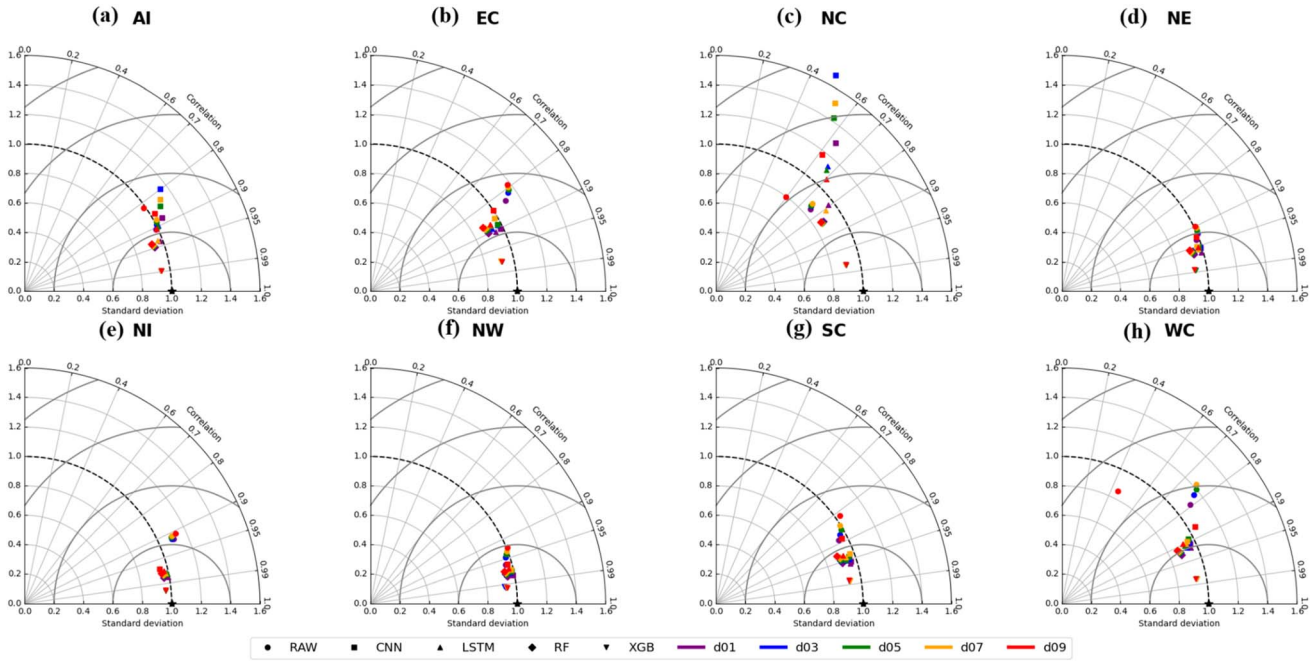

Fig. 11 Taylor diagram analysis of TMIN forecasts showing correlation, normalized standard deviation, and centered root mean square error for the raw NCUMG forecasts and ML-based bias-corrected forecasts (RF, XGB, LSTM, and CNN) at lead times Day-1, Day-3, Day-5, Day-7, and Day-9 over (a) All India (AI), (b) East Coast (EC), (c) North-Central (NC), (d) North-East (NE), (e) North India (NI), (f) North-West (NW), (g) South-Central (SC), and (h) West Coast (WC).

exceeding 30–40% and even reaching 50–60%, especially in SC. This progressive increase in RMSE improvement with lead time confirms that ML-based bias correction is most effective when systematic errors have had time to accumulate, whereas its impact is naturally limited at shorter lead times where biases are smaller. Similar results are also observed over other regions, like NE and coastal regions.

4.1.4 T_{min} RMSE improvement. For T_{min} (Fig. 9), all ML methods show a clear improvement in RMSE relative to the raw forecasts, with stronger and more spatially coherent gains than those seen for T_{max}. XGB consistently provides the largest RMSE improvements across almost all regions and lead times. Even at short lead times (Day-1), XGB shows improvements of about 25–40%, which further increase to 40–60% by Day-7 to Day-9, particularly over NC, NW, and SC regions. This indicates that XGB is effective not only at longer lead times but also when the initial bias magnitude is relatively small in the case of T_{min}.

In contrast, LSTM and RF show a clearer dependence on lead time. At Day-1, RMSE improvements from LSTM and RF are generally modest, mostly within 10–25%. As lead time increases, the improvements from both methods increase, reaching 30–45% by Day-7 and Day-9 over several stations, especially over the NC and SC regions. This behaviour is consistent with the increasing magnitude of T_{min} biases at longer lead times, where ML-based correction has greater scope to reduce errors.

The CNN method again shows the weakest performance, with limited and spatially inconsistent improvements, rarely exceeding 20–25% even at longer lead times.

This analysis shows that, unlike T_{max}, where improvements increase primarily at longer lead times, the bias correction in T_{min} using XGB is effective across all lead times, whereas RF and LSTM show progressively larger gains as lead time increases. This indicates that T_{min} errors are easier to correct using multivariate, nonlinear approaches, especially tree-based methods such as XGB.

4.1.5 Correlation and variance analysis. Fig. 10 and 11 present Taylor diagrams for T_{max} and T_{min}, respectively, across the eight regions for lead times of 1, 3, 5, 7, and 9 days, summarizing the combined behavior of correlation, normalized standard deviation, and centered RMSE (cRMSE) relative to observations. In each panel, the reference point (black star at unit standard deviation and correlation of one) denotes the observed T_{max}/T_{min}, while the distance of model points from this reference indicates the magnitude of pattern error given by cRMSE.

4.1.5.1 T_{max}. Across the AI region, the raw NCUMG forecasts lie close to the unit-standard-deviation arc, indicating realistic variability, but are farther from the reference point, showing lower correlation and higher cRMSE, with correlation decreasing with lead time. ML-based methods shift the forecasts closer to the reference by reducing cRMSE and improving pattern agreement, although with slight under-dispersion (σ < 1) indicative of mild smoothing. The ML methods show similar behaviour across lead times over the AI region.

Across NI, EC, and NE, the raw forecasts are overdispersed (σ > 1), exhibit low correlation, and have high cRMSE. ML correction increases correlation, reduces variability toward the observed level, and lowers cRMSE. Over NC and SC, the raw

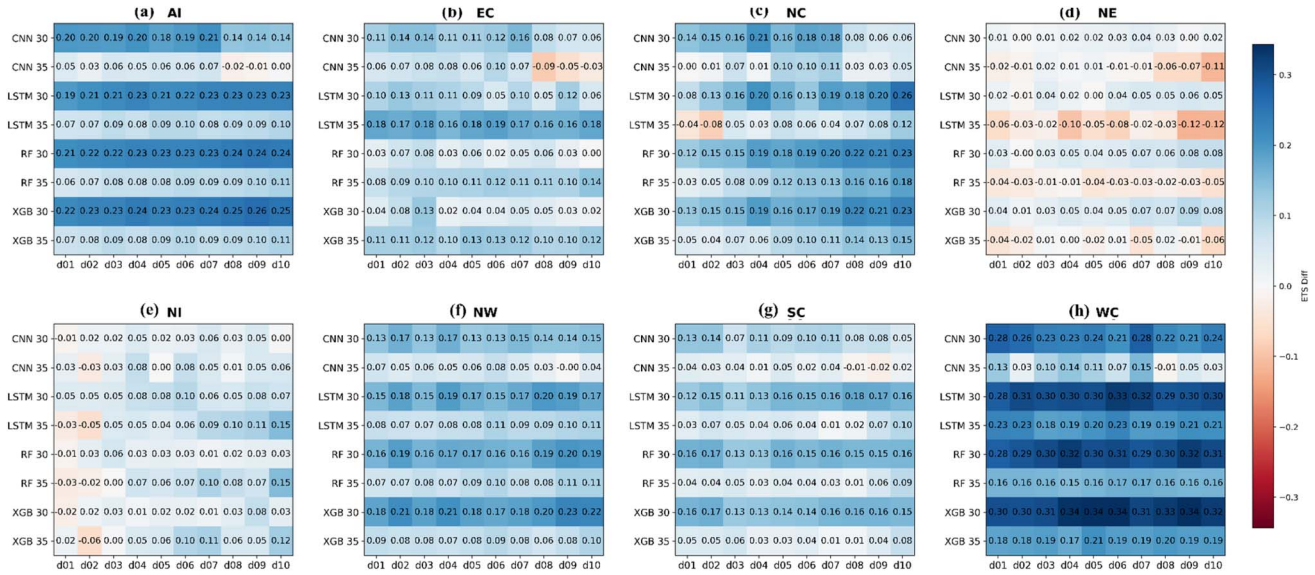

Fig. 12 Heatmap of the ETS difference between model and raw for T_{max} exceeding 30 and 35 °C during MAMJ 2023–2024. (a) All India region (AI), (b) northwest (NW), (c) north east (NE), (d) north central (NC), (e) north India (NI), (f) west coast (WC), (g) east coast (EC), and (h) south central (SC).

forecasts are under-dispersive, and ML methods improve performance by increasing variability. Over NW, where the raw forecasts already show realistic variability, ML correction improves pattern agreement by increasing correlation and reducing cRMSE. Over WC, ML post-processing improves correlation and cRMSE but introduces additional smoothing due to lower variability. However, XGB performs best in this region, achieving the highest correlation and the lowest cRMSE among all methods.

4.1.5.2 T_{min}. For the AI region, the raw NCUMG $T_{\min}$ forecasts lie close to the unit-standard-deviation arc but exhibit lower correlations, resulting in higher cRMSE. ML-based post-processing shifts the forecasts closer to the reference point by increasing correlation and reducing cRMSE, albeit with a modest reduction in variability ($\sigma < 1$), indicating slight smoothing. Among the methods, XGB achieves the highest correlation and lowest cRMSE, while CNN performs worse than the raw forecasts.

Over EC, NI, and WC, the raw forecasts show overdispersion ($\sigma > 1$), reduced correlation, and larger cRMSE. ML correction improves performance by increasing correlation, reducing variability toward unity, and lowering cRMSE. Over NE, NW,

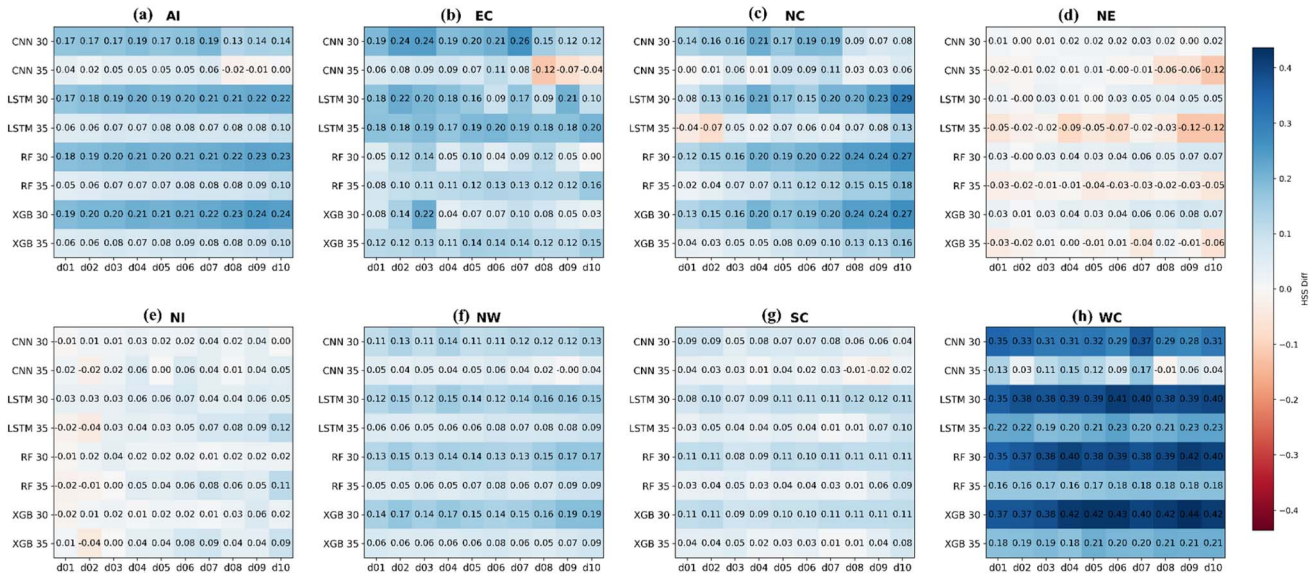

Fig. 13 Heatmap of the HSS difference between model and raw for T_{max} exceeding 30 and 35 °C during MAMJ 2023–2024. (a) All India region (AI), (b) northwest (NW), (c) north east (NE), (d) north central (NC), (e) north India (NI), (f) west coast (WC), (g) east coast (EC), and (h) south central (SC) for NCUM model TMAX forecast during 2023–2024.

and SC, the raw forecasts already exhibit variability comparable to observations, and ML-based methods primarily improve pattern agreement by increasing correlation and reducing cRMSE. In contrast, across NC, the raw forecasts are underdispersive, and most ML methods introduce overdispersion, accompanied by reduced correlation and increased cRMSE. Overall, XGB emerges as the most effective method for T_{min} bias correction across regions.

From the above analyses, it is evident that while ME and RMSE quantify the overall reduction in forecast errors, the Taylor diagrams provide additional insight into how ML-based post-processing redistributes errors between variability and pattern agreement. ML correction is most effective in regions where the raw model exhibits excessive variability or weak correlation, but may introduce over-smoothing or over-dispersion where the raw variability is already well represented. These diagnostics explain the pronounced regional dependence and the contrasting behaviour observed between T_{max} and T_{min}.

4.2 Categorical verification

ME and RMSE quantify average forecast errors but do not fully capture performance for temperature extremes. Hence, categorical verification is carried out using seasonally relevant thresholds for India ($T_{max} \geq 30$ and 35 °C during MAMJ, and $T_{min} \leq 10$ and 15 °C during DJF) to assess skill for warm days and cold nights. Results are shown as heatmaps of score differences (bias-corrected minus raw). Additionally, REV is computed to evaluate the economic benefit of bias correction for decision-making.

4.2.1 ETS for T_{max}. Fig. 12 shows the ETS difference (Model – RAW) for $T_{max} \geq 30$ °C and 35 °C from Day-1 to Day-10 across regions during March–June. Positive values indicate improved categorical skill after bias correction.

For $T_{max} \geq 30$ °C, over the AI region, RF and XGB show consistent improvements of 0.20 to 0.26, while LSTM gives 0.19 to 0.23. Over NW India, XGB shows the best improvement with differences in ETS ranging from 0.17 to 0.23, this is followed by RF 0.16 to 0.20, and LSTM 0.15 to 0.19. Over SC and NC, improvements are moderate, ranging from 0.10 to 0.20, and they increase with lead time. In contrast, NE and NI show only very small gains (0.00 to 0.08). For this threshold, large gains are seen over WC, with XGB showing an improved ETS ranging from 0.31 to 0.35 across all lead times.

For $T_{max} \geq 35$ °C, the magnitude of improvement reduces and becomes more region-dependent. Over AI, EC, NC, NW, SC, and WC, ETS gains range from +0.05 to +0.17, with NC showing a clear increase with lead time. However, over NE and NI, several methods show near-zero or negative differences (−0.02 to −0.12), indicating that bias correction does not consistently improve skill for higher T_{max} in these regions.

Across both thresholds and nearly all regions, XGB and RF consistently provide the largest ETS improvements, while CNN shows the least benefit and frequently exhibits negligible or negative changes.

4.2.2 HSS for T_{max}. Fig. 13 quantifies the HSS improvement after bias correction for $T_{max} \geq 30$ °C and ≥35 °C during March–June. For the 30 °C threshold, very large gains are evident over the WC region, where XGB, RF, and LSTM improve HSS by about 0.35 to 0.44 across lead times. Over AI, consistent gains of 0.19 to 0.24 are seen for XGB and RF, with LSTM being slightly lower (0.17 to 0.22). Over NW, improvements of 0.14 to 0.20 occur, while NC and SC show gains of 0.10 to 0.27 which increase with lead time. In contrast, NE and NI show only marginal improvements, typically ≤ 0.06. For the higher threshold of 35 °C, the magnitude of improvement decreases across all regions, with gains generally in the range 0.05 to 0.18 over AI, EC, NC, NW, SC, and WC. Over NE and NI, several methods show near-zero or negative differences (−0.03 to

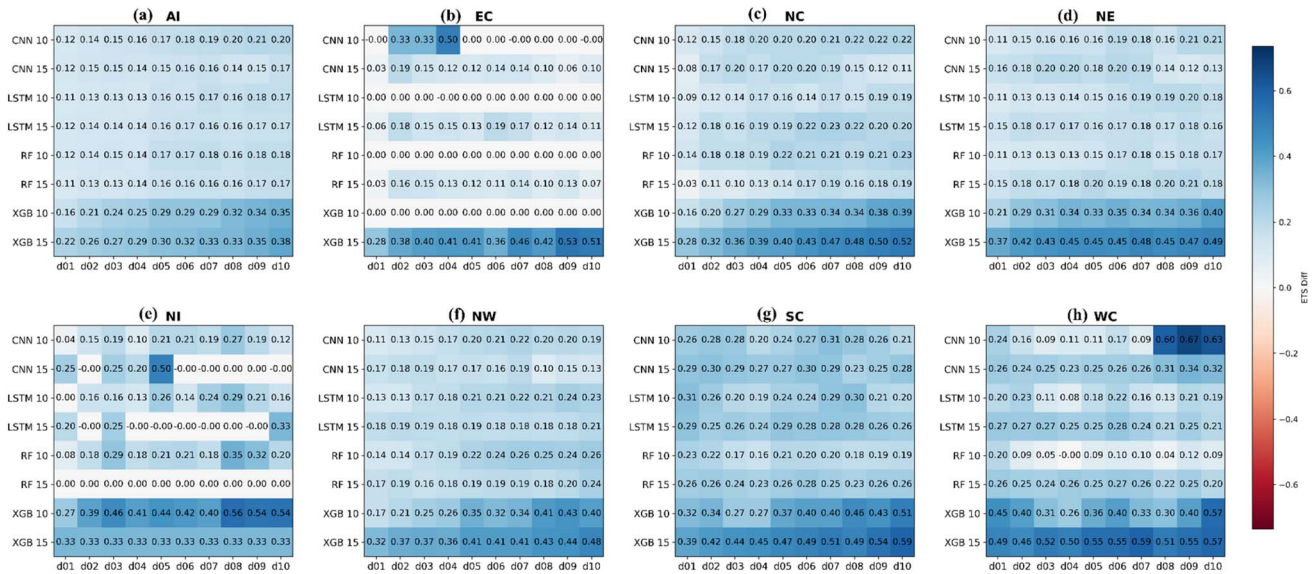

Fig. 14 Heatmap of the ETS difference between model and raw for T_{min} exceeding 10 and 15 °C during DJF 2023–2024. (a) All India region (AI), (b) northwest (NW), (c) north east (NE), (d) north central (NC), (e) north India (NI), (f) west coast (WC), (g) east coast (EC), and (h) south central (SC).

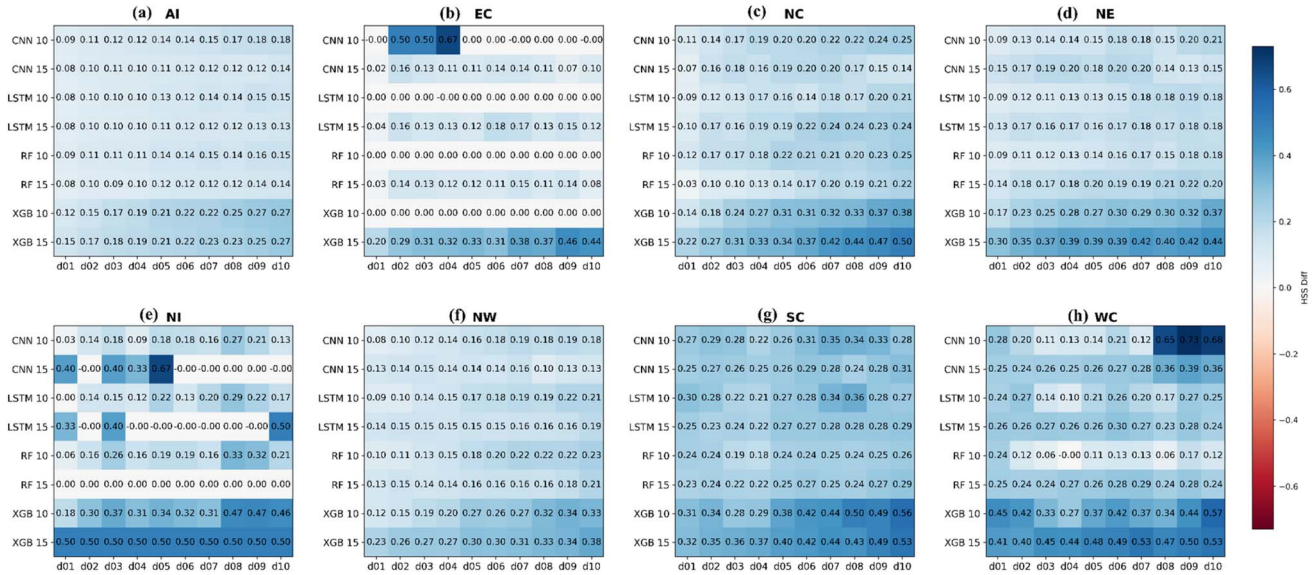

Fig. 15 Heatmap of the HSS difference between model and raw for T_{min} exceeding 10 and 15 °C during DJF 2023–2024. (a) All India region (AI), (b) northwest (NW), (c) north east (NE), (d) north central (NC), (e) north India (NI), (f) west coast (WC), (g) east coast (EC), and (h) south central (SC).

−0.12), indicating limited benefit of bias correction for more extreme heat conditions. Across both thresholds, XGB and RF provide the most consistent HSS improvement, while CNN exhibits the weakest performance.

4.2.3 ETS for T_{min}. Fig. 14 presents the ETS difference for $T_{min} \leq 10$ and 15 °C from Day-1 to Day-10 across regions during winter. The improvements are strongly threshold-dependent. For the 10 °C threshold, ETS gains are moderate across most regions, and are in the range 0.10 to 0.35, with XGB and RF

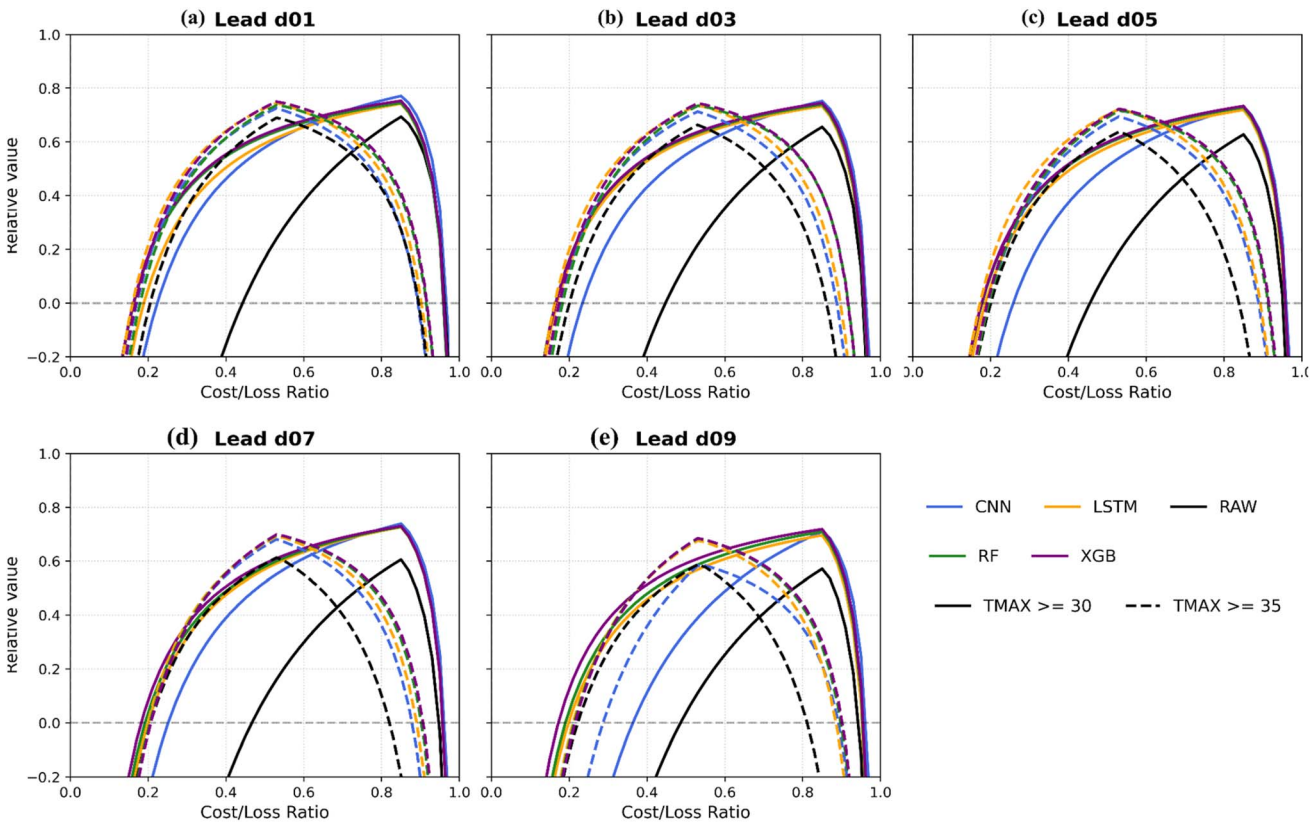

Fig. 16 REV diagram for raw and bias corrected T_{max} forecasts during MAMJ 2023–2024 for the lead times (a) Day 01, (b) Day 03, (c) Day 05 (d) Day 07 and (e) Day 09.

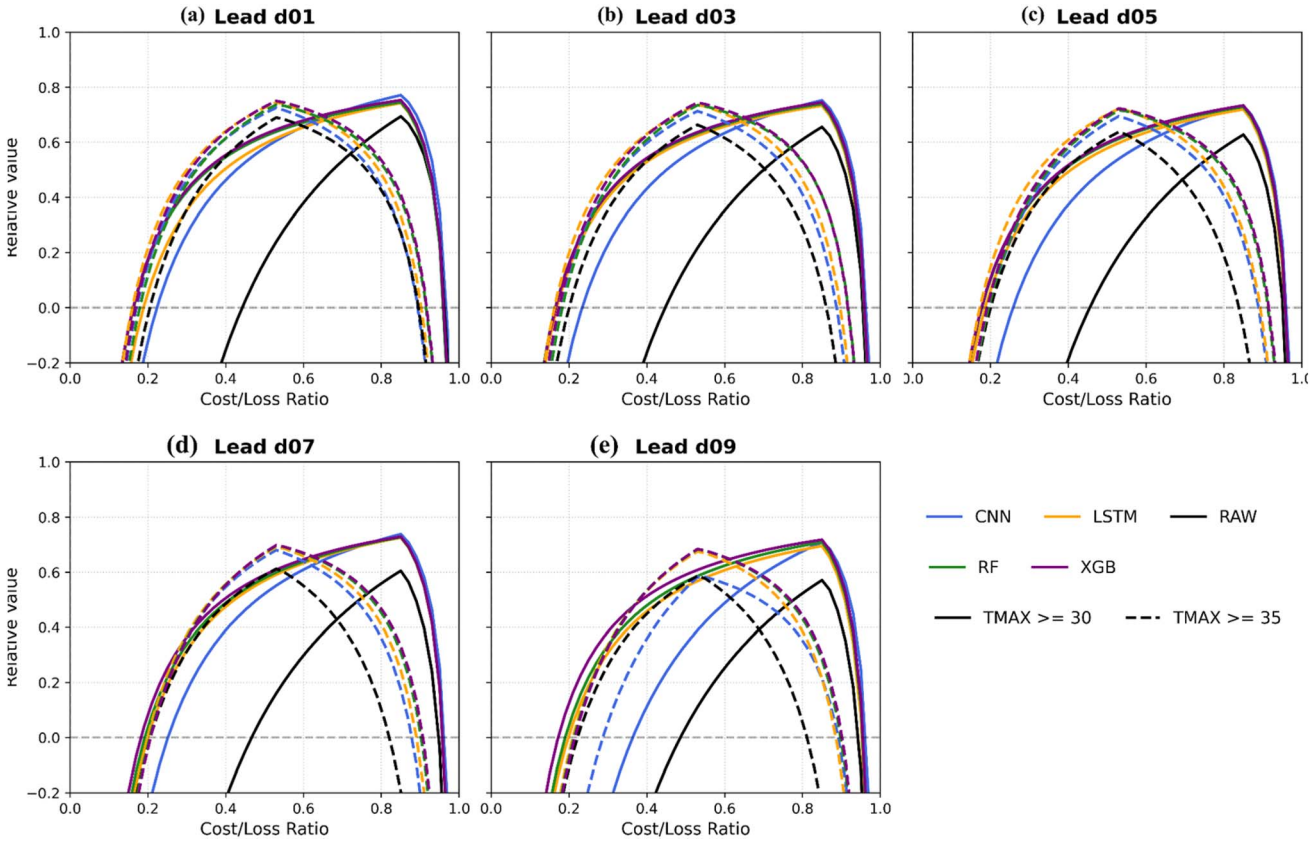

Fig. 17 REV diagram for raw and bias corrected T_{min} forecasts during DJF 2023–2024 for the lead times (a) Day 01, (b) Day 03, (c) Day 05 (d) Day 07 and (e) Day 09.

showing the most consistent improvements. Over NW and AI, gains remain modest but steady, while over SC, NC, and NE they increase gradually with lead time.

In contrast, for the 15 °C threshold, larger improvements are seen in ETS. Over WC, SC, NC, NE, EC, and NI, XGB shows gains of +0.40 to +0.60 across lead times, with RF also showing substantial improvements. These regions correspond to areas where the raw model exhibits strong systematic biases, and ML correction substantially improves the skill. Even over NW, improvements for this threshold remain consistently positive. CNN shows smaller gains compared to the other methods.

4.2.4 HSS for T_{min}. Fig. 15 shows the HSS difference which is similar to the ETS results for T_{min}. For $T_{min} \leq 10$ °C, HSS improvements are moderate, generally between 0.10 and 0.35, with XGB and RF providing consistent gains across regions and lead times. For $T_{min} \leq 15$ °C, the HSS gains become highly pronounced. Over WC, SC, NC, NE, EC, and NI, XGB frequently achieves improvements of 0.45 to 0.65, indicating better performance of the model in predicting higher T_{min} during winter season. RF shows similar but slightly smaller gains, while CNN provides the least improvement.

These results show that ML-based bias correction, particularly using XGB, is effective in improving the model's ability to correctly identify winter cold-night conditions across India, especially for the more climatologically frequent threshold (≤15 °C).

These sections show that ML-based bias correction, especially using XGB (and RF), substantially improves the model's ability to correctly identify warm and cold temperature events, as seen by consistent positive ETS and HSS differences across most regions and lead times. The improvements are greatest in regions where the raw model has poor categorical skill, particularly for winter T_{min} and summer T_{max} over WC, EC, NC, and SC, while gains are smaller in regions where the raw variability and skill are already reasonable (NE and NI).

4.2.5 Relative economic value. Bias correction may lead to a reduced economic value of forecasts, particularly when it introduces over-smoothing of the data. In such cases, extreme events may be inadequately represented, leading either to unnecessary precautionary actions (incurring avoidable costs) or to missed events (resulting in high losses). This is closely linked to the discrimination ability of the model, *i.e.*, its capacity to distinguish between extreme and non-extreme events.[67]

4.2.5.1 REV for T_{max}. Across all lead times, the ML-corrected forecasts, particularly from XGB and RF, consistently provide higher REV (Fig. 16) than the raw model for both thresholds ($T_{max} \geq 30$ °C and ≥ 35 °C). For the 30 °C threshold, the raw forecast shows positive value only over a narrower range of cost/loss ratios (≈0.4–1.0), whereas bias-corrected forecasts yield positive values across a much wider range (≈0.2–1.0). A similar pattern is observed for the 35 °C threshold, where bias

correction not only expands the range of cost/loss ratios for which the forecasts are decision-useful, but also increases the peak economic value. As lead time increases (Day-01 to Day-09), the value of the raw model decreases, whereas the ML-based forecasts retain meaningful economic value, demonstrating that post-processing not only reduces statistical error but also extends the relevance of the models for decision-making even at longer lead times. It is seen that XGB and RF yield the highest REV values at all lead times, followed by LSTM. CNN performs worse than the other ML models.

4.2.5.2 REV for T_{min}. For T_{min} (Fig. 17), the value diagrams show a much stronger separation between the raw and ML-corrected forecasts than for T_{max}. The raw NCUMG forecasts provide useful economic value only over a limited range of cost/loss ratios ($\leq$0.8 for Day-01 and Day-03 lead times), and this range narrows further with lead time, along with the peak value, particularly for $T_{min} \leq 10$ °C. In contrast, forecasts obtained by using XGB and RF substantially increase both the maximum REV and the range of cost/loss ratios over which the forecasts remain useful for decision making across all lead times.

XGB shows the highest REV as compared to all models for both $T_{min} \leq 10$ °C and 15 °C. A notable feature is the flatness of the XGB curve, indicating that the forecast retains high economic value regardless of user cost/loss preferences. This is associated with strong discrimination and low false alarm rates, indicating that the model correctly identifies cold-night events without frequently triggering unnecessary actions.

5. Conclusions

In this study, we have sought to correct the systematic biases in NCMRWF Unified Model (NCUMG) T_{max} and T_{min} forecasts across India using machine learning (ML)-based bias-correction frameworks. Using a comprehensive station network and multivariate predictor set derived from NCUMG forecasts, several ML methods, including Random Forest (RF), eXtreme Gradient Boost (XGB), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN), were evaluated for their ability to correct forecast errors across different climatic regions and lead times. The effectiveness of these methods was assessed using continuous, categorical, and decision-oriented verification metrics, enabling a comprehensive evaluation of how bias correction improves not only forecast accuracy but also forecast usability. The salient conclusions drawn from this study are presented below:

SHAP analysis confirms that the predictors driving bias correction are physically consistent with known daytime and nighttime temperature processes, lending interpretability to the ML framework and providing confidence in the physical basis of the corrections applied. ML-based bias correction substantially reduces mean error and RMSE for both T_{max} and T_{min} across all regions and lead times, with improvements becoming more pronounced at longer lead times where raw model biases grow. Among the four methods evaluated, XGB consistently emerges as the best performing method across regions and lead times, particularly for T_{min}, followed by RF, LSTM, and CNN respectively. Analysis based on Taylor diagrams further shows that ML correction improves not only error magnitude but also corrects both over-dispersion and under-dispersion present in the raw NCUMG forecasts. Categorical verification using ETS and HSS shows large positive skill improvements across most regions and lead times, particularly for winter T_{min} and summer T_{max} over WC, EC, NC, and SC where the raw model has weak skill, while regions such as NE and NI show comparatively smaller categorical improvements as the raw model already represents temperature variability reasonably well in these areas. REV analysis confirms that ML-corrected forecasts remain useful over a much wider range of cost/loss ratios and retain economic value at longer lead times, unlike the raw model whose value diminishes rapidly beyond Day-3. The results clearly show that ML-based bias correction, with XGB emerging as the most consistently skilful method, delivers meaningful improvements in accuracy, categorical skill, and economic value across diverse climatic regions and lead times, offering a practically viable and immediately deployable enhancement to operational T_{max} and T_{min} predictions over India. It is worth noting that the LSTM and CNN architectures as implemented here operate on a single predictor vector per lead time at each station independently and therefore do not fully exploit their temporal sequencing and spatial feature extraction capabilities. Earlier studies and our own recent work have shown that under comparable station level tabular predictor conditions, tree-based methods perform at least as well as neural network approaches, a finding our results are entirely consistent with. The real promise of LSTM and CNN for this problem lies in autoregressive implementations combining past observations with forecast sequences and spatially aware CNN frameworks incorporating geographic information such as orography and land cover. Work is already underway at NCMRWF to develop an advanced ML framework combining CNN with statistical methods for gridded T_{max} and T_{min} bias correction from operational NWP models, building directly on the baseline this study has established.

This study demonstrates that ML-based bias correction, particularly using XGB, enhances forecast accuracy, categorical skill, and economic value thus providing an improved T_{max} and T_{min} prediction across India. It is worth noting, however, that LSTM and CNN as implemented here operate on a single predictor vector per lead time at each station independently, and therefore do not fully exploit temporal sequencing and spatial feature extraction. The potential of LSTM and CNN can be fully realised through temporally aware autoregressive implementations and spatially aware frameworks incorporating geographic fields such as orography and land–sea mask. Work is ongoing at NCMRWF to develop a hybrid ML framework combining CNN with statistical methods for gridded T_{max} and T_{min} bias correction from operational NWP.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data can be made available on request.

Acknowledgements

During the preparation of this work the author(s) used Grammarly for language and grammar improvement. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

1 P. Rohini, M. Rajeevan and A. K. Srivastava, On the Variability and Increasing Trends of Heat Waves over India, *Sci. Rep.*, 2016, **6**, 26153, DOI: 10.1038/srep26153.
2 D. K. Panda, A. AghaKouchak and S. K. Ambast, Increasing heat waves and warm spells in India, observed from a multiaspect framework, *J. Geophys. Res.*, 2017, **122**(7), 3837–3858, DOI: 10.1002/2016JD026292.
3 T. N. Palmer, The economic value of ensemble forecasts as a tool for risk assessment: From days to decades, *Q. J. R. Metereol. Soc.*, 2002, **128**(581), 747–774, DOI: 10.1256/0035900021643593.
4 E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, Cambridge, vol. 6, 2003.
5 M. Leutbecher and T. N. Palmer, Ensemble forecasting, *J. Comput. Phys.*, 2008, **227**(7), 3515–3539, DOI: 10.1016/j.jcp.2007.02.014.
6 P. Bauer, A. Thorpe and G. Brunet, The quiet revolution of numerical weather prediction, *Nature*, 2015, **525**(7567), 47–55, DOI: 10.1038/nature14956.
7 D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd edn, Academic Press, Oxford, 2011.
8 S. Watanabe, S. Kanae, S. Seto, P. J. -F. Yeh, Y. Hirabayashi and T. Oki, Intercomparison of bias-correction methods for monthly temperature and precipitation simulated by multiple climate models, *J. Geophys. Res. Atmos.*, 2012, **117**(D23), 1–13, DOI: 10.1029/2012JD018192.
9 X. Yang, E. F. Wood, J. Sheffield, L. Ren, M. Zhang and Y. Wang, Bias Correction of Historical and Future Simulations of Precipitation and Temperature for China from CMIP5 Models, *J. Hydrometeorol.*, 2018, **19**(3), 609–623, DOI: 10.1175/JHM-D-17-0180.1.
10 V. M. Garibay, M. W. Gitau, N. Kiggundu, D. Moriasi and F. Mishili, Evaluation of Reanalysis Precipitation Data and Potential Bias Correction Methods for Use in Data-Scarce Areas, *Water Resour. Manag.*, 2021, **35**(5), 1587–1602, DOI: 10.1007/s11269-021-02804-8.
11 M. R. Haider, M. Peña and E. Anagnostou, Bias Correction of Mixed Distributions of Temperature with Strong Diurnal Signal, *Weather Forecast.*, 2022, **37**(4), 495–509, DOI: 10.1175/WAF-D-21-0108.1.
12 P. Dhawan, D. Dalla Torre, M. Niazkar, K. Kaffas, M. Larcher, M. Righetti, *et al.*, A comprehensive comparison of bias correction methods in climate model simulations: Application on ERA5-Land across different temporal resolutions, *Heliyon*, 2024, **10**(23), 1–16, DOI: 10.1016/j.heliyon.2024.e40352.
13 V. R. Durai and R. Bhradwaj, Evaluation of statistical bias correction methods for numerical weather prediction model forecasts of maximum and minimum temperatures, *Nat. Hazards*, 2014, **73**(3), 1229–1254, DOI: 10.1007/s11069-014-1136-1.
14 H. Singh, A. Dube, S. Kumar and R. Ashrit, Bias correction of maximum temperature forecasts over India during March–May 2017, *J. Earth Syst. Sci.*, 2020, **129**(13), DOI: 10.1007/s12040-019-1291-6.
15 V. K. Valappil, M. Temimi, M. Weston, R. Fonseca, N. R. Nelli, M. Thota, *et al.*, Assessing Bias Correction Methods in Support of Operational Weather Forecast in Arid Environment, *Asia Pac. J. Atmos. Sci.*, 2020, **56**(3), 333–347, DOI: 10.1007/s13143-019-00139-4.
16 D. Dutta and R. K. Bhattacharjya, A statistical bias correction technique for global climate model predicted near-surface temperature in India using the generalized regression neural network, *J. Water Clim. Change*, 2022, **13**(2), 854–871, DOI: 10.2166/wcc.2022.214.
17 D. S. Wilks, Statistical Methods in the Atmospheric Sciences: An Introduction, *1st Ed. 59 of Int.Geophysics*, Elsevier, 1995, pp. 1–467.
18 R. Hagedorn, F. J. Doblas-Reyes and T. N. Palmer, The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept, *Tellus Dyn. Meteorol. Oceanogr.*, 2005, **57**(3), 219, DOI: 10.3402/tellusa.v57i3.14657.
19 M. Niazkar, R. Piraei, A. Menapace, P. Dhawan, D. D. Torre, M. Larcher, *et al.*, Bias correction of ERA5-Land temperature data using standalone and ensemble machine learning models: a case of northern Italy, *J. Water Clim. Change*, 2024, **15**(1), 271–283, DOI: 10.2166/wcc.2023.669.
20 S. Guo, Y. Yang, F. Zhang, J. Wang and Y. Cheng, Study on bias correction method of ECMWF surface variable forecasts based on deep learning, *Renew. Energy*, 2025, 239, DOI: 10.1016/j.renene.2024.122132.
21 H. Singh, A. Dube, P. K. Srivastava, R. Ashrit, J. P. George and V. S. Prasad, Improving the Indian Monsoon Data Assimilation and Analysis regional reanalysis and maximum and minimum temperatures over India using macine-learning techniques, *Q. J. R. Metereol. Soc.*, 2026, **152**(776), e70093, DOI: 10.1002/qj.70093.
22 S. Veldkamp, K. Whan, S. Dirksen and M. Schmeits, Statistical Postprocessing of Wind Speed Forecasts Using Convolutional Neural Networks, *Mon. Weather Rev.*, 2021, **149**(4), 1141–1152, DOI: 10.1175/MWR-D-20-0219.1.
23 D. Cho, C. Yoo, J. Im and D. H. Cha, Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas, *Earth Space Sci.*, 2020, **7**(4), 1–18, DOI: 10.1029/2019EA000740.
24 R. Roberts, A. Wong, S. Jenkins, A. Neher, C. Sutton, P. O'Meara, *et al.*, Mental health and well-being impacts of COVID-19 on rural paramedics, police, community nurses and child protection workers, *Aust. J. Rural Health*, 2021, **29**(5), 753–767, DOI: 10.1111/ajr.12804.

25 S. Kelkar and K. Dairaku, Investigation of Uncertainties in Multi-variable Bias Adjustment in Multi-model Ensemble, *Proc. IAHS*, 2024, **386**, 55–60, DOI: 10.5194/piahs-386-55-2024.
26 K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu and Q. Tian, Accurate medium-range global weather forecasting with 3D neural networks, *Nature*, 2023, **619**(7970), 533–538.
27 R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed and P. Battagilla, Learning skillful medium-range global weather forecasting, *Science*, 2023, **382**(6677), 1416–1421, DOI: 10.1126/science.adi23.
28 R. Stephan, S. Hoyer, A. Merose, I. Langmore, P. Battagilla, T. Russel, A. Sanchex-Gonzalez, V. Yang, R. Curver, S. Agrawal, M. Chantry, B. Z. Bouallegue, P. Dueben, C. Bromberg, J. Sisk, L. Barrington, A. Bell and F. Sha, WeatherBench 2: A benchmark for the next generation of data-driven global weather models, *J. Adv. Model. Earth Syst.*, 2024, **16**(6), e2023MS004019, DOI: 10.1029/2023MS004019.
29 B. Trotta, R. Johnson, C. de Burgh-Day, D. Hudson, E. Abellan, J. Canvin, *et al.*, Statistical Postprocessing Yields Accurate Probabilistic Forecasts from Artificial Intelligence Weather Models, *Artif. Intell. Earth Syst.*, 2025, **4**(4), 1–17, DOI: 10.1175/AIES-D-25-0037.1.
30 L. Breiman, Random Forests, *Mach. Learn.*, 2001, **45**(1), 5–32, DOI: 10.1023/A:1010933404324.
31 J. Wang, Q. Cheng and Y. Dong, An XGBoost-based multivariate deep learning framework for stock index futures price forecasting, *Kybernetes*, 2023, **52**(10), 4158–4177, DOI: 10.1108/K-12-2021-1289.
32 S. Kumar, A. Dube, R. Ashrit and A. K. Mitra, A Machine Learning (ML)-Based Approach to Improve Tropical Cyclone Intensity Prediction of NCMRWF Ensemble Prediction System, *Pure Appl. Geophys.*, 2023, **180**(1), 261–275, DOI: 10.1007/s00024-022-03206-6.
33 P. Feng, R. Wang, J. Sun, W. Yan, P. Chi and X. Luo, An interpretable ensemble machine-learning workflow for permeability predictions in tight sandstone reservoirs using logging data, *Geophysics*, 2024, **89**(5), MR265–MR280, DOI: 10.1190/geo2023-0657.1.
34 Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner and W. Collins, Application of deep convolutional neural networks for detecting extreme weather in climate datasets, *arXiv*, 2016, preprint, arXiv:1605.01156, DOI: 10.48550/arXiv.1605.01156.
35 S. C. M. Sharma, B. Kumar, A. Mitra and S. K. Saha, Deep learning-based bias correction of ISMR simulated by GCM, *Atmos. Res.*, 2024, 309, DOI: 10.1016/j.atmosres.2024.107589.
36 S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, *Neural Comput.*, 1997, **9**, 1735–1780.
37 S. N. H. Bukhari and K. A. Ogudo, Forecasting temperature and rainfall using deep learning for the challenging climates of Northern India, *PeerJ Comput. Sci.*, 2025, **11**, e3012, DOI: 10.7717/peerj-cs.3012.
38 S. K. Dash, R. K. Jenamani, S. R. Kalsi and S. K. Panda, Some evidence of climate change in twentieth-century India, *Clim. Change*, 2007, **85**(3–4), 299–321, DOI: 10.1007/s10584-007-9305-9.
39 S. K. Dash and A. Mamgain, Changes in the Frequency of Different Categories of Temperature Extremes in India, *J. Appl. Meteorol. Climatol.*, 2011, **50**(9), 1842–1858, DOI: 10.1175/2011JAMC2687.1.
40 S. K. Dash, In. *Extreme Temperature Regions in India*. 2016, pp. 55–73, DOI: 10.1007/978-3-319-23684-1_4.
41 *Imds_Vision_*, 2047.
42 D. R. Kothawale, N. R. Deshpande and R. K. Kolli, Long Term Temperature Trends at Major, Medium, Small Cities and Hill Stations in India during the Period 1901-2013, *Am. J. Clim. Change*, 2016, **05**(03), 383–398, DOI: 10.4236/ajcc.2016.53029.
43 A. K. Srivastava, M. Rajeevan and S. R. Kshirsagar, Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region, *Atmos. Sci. Lett.*, 2009, **10**(4), 249–254, DOI: 10.1002/asl.232.
44 J. Thuburn, ENDGame: The New Dynamical Core of the Met Office Weather and Climate Prediction Model, In. *UK Success Stories in Industrial Mathematics*, Springer International Publishing, Cham, 2016, pp. 27–33, DOI: 10.1007/978-3-319-25454-8_4.
45 D. Barker. Data assimilation-progress and plans, In. *MOSAC-16*. MOSDAC, 2011.
46 S. Kumar, G. George, B. P. J., M. T. Bushair, S. Indira Rani and J. P. George, *et al.*, NCUM Global DA System: Highlights of the 2021 upgrade, Technical report, 2021, vol. 5, pp. 1–28, nmrf.tr.5.2021, https://nwp.ncmrwf.gov.in/publications/tr/10.64349/nmrf.tr.5.2021.
47 S. Kumar, M. T. Bushair, B. P. J., A. Lodh, P. Sharma, G. George, S. Indira Rani, J. P. George, A. Jayakumar, S. Mohandas, S. Kumar, K. Sharma, S. Karunasagar and E. N. Rajagopal, *et al.*, NCUM Global NWP System: Version 6 (NCUM-G:V6), Technical report, 2020, vol. 6pp. 1–32, nmrf.tr.6.2020, https://nwp.ncmrwf.gov.in/publications/tr/10.64349/nmrf.tr.6.2020.
48 S. Kumar, A. Jayakumar, M. T. Bushair, B. P. J., G. George, A. Lodh, S. Indira Rani, S. Mohandas, J. P. George and E. N. Rajagopal, *et al.*, Implementation of New High Resolution NCUM Analysis-Forecast System in Mihir HPCS, Technical report, 2018, vol. 1, pp. 1–17, nmrf.tr.1.2018, https://nwp.ncmrwf.gov.in/publications/tr/10.64349/nmrf.tr.1.2018.
49 R. Ashrit, R. S. Indira, S. Kumar, S. Karunasagar, T. Arulalan, T. Francis, A. Routray, S. I. Laskar, S. Mahmood, P. Jermey, A. Maycock, R. Renshaw, J. P. George and E. N. Rajagopal, IMDAA Regional Reanalysis: Performance Evaluation During Indian Summer Monsoon Season, *J. Geophys. Res. Atmos.*, 2020, **125**(2), 1–26, DOI: 10.1029/2019JD030973.
50 S. I. Rani, T. Arulalan, J. P. George, E. N. Rajagopal, R. Renshaw, A. Maycock, M. D. Barker and M Rajeevan, IMDAA: High Resolution Satellite-era Reanalysis for the Indian Monsoon Region, *J. Clim.*, 2021, 5109–5133, DOI: 10.1175/JCLI-D-20-0412.1.

51 E. J. Barton, C. M. Taylor, A. K. Mitra and A. Jayakumar, Systematic daytime increases in atmospheric biases linked to dry soils in irrigated areas in Indian operational forecasts, *Atmos. Sci. Lett.*, 2023, **24**(9), 1–11, DOI: 10.1002/asl.1172.

52 M. A. Babyak, What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models, *Psychosom. Med.*, 2004, **66**(3), 411–421, DOI: 10.1097/01.psy.0000127692.23278.a9.

53 J. Lever, M. Krzywinski and N. Altman, Model selection and overfitting, *Nat. Methods*, 2016, **13**(9), 703–704, DOI: 10.1038/nmeth.3968.

54 J. Y. L. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z. W. Hong, *et al.*, Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review, *Mathematics*, 2022, **10**(8), 1283, DOI: 10.3390/math10081283.

55 N. Shrestha, Detecting Multicollinearity in Regression Analysis, *Am. J. Appl. Math. Stat.*, 2020, **8**(2), 39–42, DOI: 10.12691/ajams-8-2-1.

56 S. Kumar, A. Dube, R. Ashrit and A. K. Mitra, A Machine Learning (ML)-Based Approach to Improve Tropical Cyclone Intensity Prediction of NCMRWF Ensemble Prediction System, *Pure Appl. Geophys.*, 2023, **180**(1), 261–275, DOI: 10.1007/s00024-022-03206-6.

57 L. Breiman, Random Forests, *Mach. Learn.*, 2001, **45**(1), 5–32, DOI: 10.1023/A:1010933404324.

58 J. Wang, Q. Cheng and Y. Dong, An XGBoost-based multivariate deep learning framework for stock index futures price forecasting, *Kybernetes*, 2023, **52**(10), 4158–4177, DOI: 10.1108/K-12-2021-1289.

59 W. Liu, L. Deng, H. Li, X. Li, C. Shi, N. Meng, *et al.*, Experimental study and machine learning on the maximum temperature beneath tunnel ceiling induced by adjacent tandem fires in longitudinally ventilated tunnel, *Int. J. Therm. Sci.*, 2023, **187**, 108169, DOI: 10.1016/j.ijthermalsci.2023.108169.

60 J. T. Schaefer, The Critical Success Index as an Indicator of Warning Skill, *Weather Forecast.*, 1990, **5**(4), 570–575, DOI: 10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.

61 P. Heidke, Berechnung der erfolges und der gute der windstarkevorhersagen im sturmwarnungdienst, *Geogr. Ann.*, 1926, **8**, 301–349.

62 R. W. Katz and A. H. Murphy, *Economic Value of Weather and Climate Forecasts*. Cambridge University Press., 1997.

63 H. R. Marzban and S. M. Hoseini, Solution of Nonlinear Volterra-Fredholm Integrodifferential Equations via Hybrid of Block-Pulse Functions and Lagrange Interpolating Polynomials, *Adv. Numer. Anal.*, 2012, **2012**, 1–14, DOI: 10.1155/2012/868279.

64 Z. He, Y. Yang, R. Fang, S. Zhou, W. Zhao, Y. Bai, *et al.*, Integration of shapley additive explanations with random forest model for quantitative precipitation estimation of mesoscale convective systems, *Front. Environ. Sci.*, 2023, **10**, 1–15, DOI: 10.3389/fenvs.2022.1057081.

65 Z. Song, S. Cao and H. Yang, An interpretable framework for modeling global solar radiation using tree-based ensemble machine learning and Shapley additive explanations methods, *Appl. Energy*, 2024, **364**, 123238, DOI: 10.1016/j.apenergy.2024.123238.

66 C. Xiao, A. Duan, Y. Tang, B. Tang, Q. Wang and X. Yang, Machine learning prediction of summer extreme precipitation days in the middle and lower Yangtze River with SHAP explanation, *Atmos. Res.*, 2026, **330**, 108614, DOI: 10.1016/j.atmosres.2025.108614.

67 S. J. Mason, Understanding forecast verification statistics, *Front. Environ. Sci.*, 2008, **15**(1), 31–40, DOI: 10.1002/met.51.