Improving Tmax Predictions with Multimodel Machine Learning Approaches

Harvir Singh*[1,2]*, *Anumeha Dube*[1], *Raghavendra Ashrit*[1] *Prashant Kumar Srivastava*[2] *and VS Prasad*[1]
[1]*National Centre for Medium-Range Weather Forecasting, Ministry of Earth Sciences, Noida, India*
[2]*Institute of Environmental and Sustainable Development, Banaras Hindu University, Varanasi, India*
****Corresponding Author: harvir.ncmrwf@gmail.com***
https://orcid.org/0000-0003-1355-7628

Introduction

Extreme weather events such as heavy rainfall, flash floods, and heat waves have increased in recent years, posing significant threats to lives and property. Accurate weather forecasting is essential for disaster preparedness, particularly for predicting temperature, a critical factor in agriculture, water management, and human health. Modern weather forecasting has evolved from station-based to fine-gridded grid forecasts, emphasizing the need for precision. Numerical Weather Prediction (NWP) models play a central role, but limitations in modeling complex terrains and local phenomena necessitate post-processing techniques like Model Output Statistics (MOS) and Kalman filtering to correct biases.

Traditional methods often struggle with nonlinear atmospheric dynamics, prompting researchers to leverage machine learning (ML) for bias correction. ML models excel at processing multidimensional data and identifying nonlinear relationships. Studies show advanced methods like Extreme Gradient Boosting and deep learning architectures—e.g., recurrent neural networks (RNNs) and convolutional neural networks (CNNs)—effectively improve forecast accuracy. For example, CNNs have been used to enhance grid-based corrections, while RNNs address spatio-temporal complexities.

Despite significant progress, multimodel ML-based approaches remain underutilized. This study introduces a novel ML-driven multimodel tool that synthesizes forecasts, reduces systematic errors, and aids forecasters in improving reliability and precision in temperature predictions.

Methodology

To improve forecasting maximum temperatures (Tmax) at different stations, four machine learning (ML) algorithms—Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGB), and Multiple Linear Regression (MLR)—were applied to correct biases in the ensemble mean of the GEFS model. These biases stem from issues in model physics, initial conditions, and ensemble configuration.

ML Techniques Applied:

1. **SVM**: Utilizes kernel functions to handle nonlinear data relationships by mapping them into higher-dimensional spaces. Implemented with the 'e1071' R package, it minimizes hinge loss and constructs a hyperplane fitting the data with a defined margin.
2. **RF**: Combines decision trees through bagging to enhance accuracy and robustness. Implemented using the 'randomForest' package, RF tunes parameters like the number of trees (ntree) and variables (mtry) to optimize performance.
3. **XGB**: Employs gradient boosting to iteratively reduce errors by adding weighted decision trees, penalizing complexity to avoid overfitting. Tuned using the 'xgboost' package with parameters like max.depth and nrounds.
4. **MLR**: Models linear relationships between dependent and explanatory variables, testing variable significance and predicting outcomes.

The historical data was split into training and testing sets, with algorithms trained to minimize errors and evaluated against the ensemble mean. Bias corrections extended beyond the ensemble mean to individual ensemble members while preserving the ensemble spread. This was achieved by adjusting raw ensemble members based on their differences from both raw and corrected means. This comprehensive approach enhances Tmax forecasts and supports operational implementation.

Results

The performance of machine learning (ML) techniques for predicting maximum temperature (Tmax) using the Global Ensemble Forecast System (GEFS) demonstrates notable effectiveness across the Indian region. Among the techniques, Support Vector Machine (SVM) and Multiple Linear Regression (MLR) stand out, delivering favorable results with Root Mean Square Error (RMSE) reductions ranging between 23% and 45% across various stations. The most remarkable improvement is observed at Gorakhpur Station, where SVM and MLR reduce the RMSE of the GEFS mean Tmax forecast by 43.03% for Day-1, 42.00% for Day-3, and 36.60% for Day-5, signifying their reliability in temperature prediction over short- to medium-term forecasts.

In contrast, Random Forest (RF) techniques exhibit the highest RMSE values across all stations, underperforming in comparison to SVM and MLR. These findings suggest that while ML techniques overall enhance Tmax forecasting, the choice of algorithm significantly impacts the accuracy, with RF facing challenges in this context.

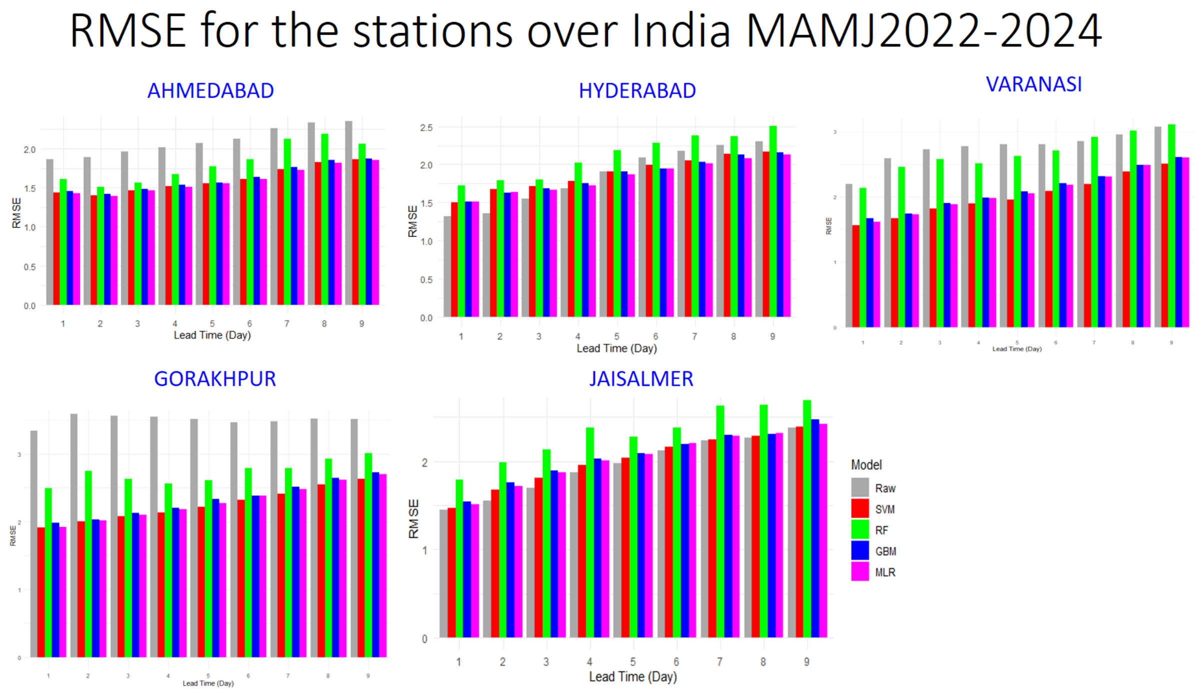

Figure 1. Ensemble Tmax mean RMSE for the stations, Ahmedabad, Hyderabad, Gorakhpur, Jaisalmer and Varanasi for the model GEFS and ML models SVM, RF, GBM, and MLR respectively.

References:

Cho, D., Yoo, C., Im, J., & Cha, D.-H. (2020). Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. Earth and Space Science, 7, e2019EA000740. https://doi.org/10.1029/2019EA000740