The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 2026

Volume 40, Pages 391-396

ICBAST 2026: International Conference on Basic Sciences and Technology

A Machine Learning Framework for Long-term Precipitation Prediction Using Gradient Boosting and Ensemble Techniques

Nexhmije Miftari
Mother Teresa University

Valon Ahmeti
Mother Teresa University

Majlinda Axhiu
Mother Teresa University

Abstract: Long-term precipitation prediction remains a challenging task due to the nonlinear and dynamic interactions among meteorological variables, particularly in regions with complex climatic conditions. This study proposes a Machine Learning framework based on XGBoost combined with temporal feature engineering for predicting rainfall, snowfall, and total precipitation in North Macedonia. Given the limitations of traditional numerical weather prediction models and the lack of region-specific studies focused on the Balkan climate, this research investigates whether advanced supervised learning algorithms can improve the accuracy and stability of precipitation forecasting.A comprehensive dataset covering the period from 2000 to 2025 was obtained from Visual Crossing, followed by extensive preprocessing and feature engineering. The engineered features include lag variables, rolling statistics, seasonal indicators, and interaction terms to better capture temporal dependencies and complex relationships in the data.Three predictive models were developed: a rainfall classifier, a snowfall classifier, and a general precipitation classifier. The primary algorithm employed was XGBoost, while the snowfall prediction model utilized an ensemble stacking approach combining XGBoost, Random Forest, and LightGBM to address class imbalance and enhance model robustness. Model validation was conducted using a TimeSeriesSplit strategy with temporal gaps to prevent data leakage.The results demonstrate significant performance improvements using gradient boosting-based methods. The snowfall model achieved an accuracy ranging from 90% to 97%, rainfall prediction reached 81% to 84%, and the general precipitation model achieved 80% to 81% accuracy. Feature importance analysis identified temperature, humidity, atmospheric pressure patterns, and cloud cover as the most influential predictors. These findings confirm that machine learning provides a reliable and effective approach for long-term precipitation forecasting in the region, with potential applications in agriculture, transportation, and environmental planning.

Keywords: XGBoost, Random Forest, LightGBM, Precipitation prediction, Predictive machine learning.

Introduction

Accurate forecasting of precipitation is essential for many sectors, including agriculture, transportation, and urban development. Long-range predictions continue to be problematic due to the intricate relationships among atmospheric factors, as highlighted by Doblas-Reyes et al. (2013). This difficulty is particularly noticeable in North Macedonia, where the combination of continental and Mediterranean climates generates a considerable amount of change in local weather (Schaer et al., 2016).

Traditional forecasting methods including numerical forecast models are based on physical equations and large scale simulations. While these methods are essential, they require a lot of computing power and can be sensitive

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

to the initial conditions (Hoffman & Kalnay, 1983). Furthermore, statistical approaches like linear regression or ARIMA frequently have trouble with the nonlinear dynamics found in meteorological measurements. In this context, machine learning is emerging as a promising option. By utilizing past data, it can determine complicated patterns without being dependent on defined physical principles (Agarwal & Kumar, 1983). XGBoost, among the different options, has shown excellent results with structured data because of its ability to model nonlinear interactions (Chen & Guestrin, 2016).

Using long-term meteorological data from Skopje this study aims to create a solid ML framework for predicting rainfall, snowfall and total precipitation. The focus is in particular building multiple predictive models suited to different types of precipitation using feature engineering techniques to capture temporal dependencies assessing model performance with time-aware validation strategies and analyzing feature importance to gain insights.

Method

Dataset and Preprocessing

The data is a collection of daily weather readings from 2000 to 2025, a total of over 9,000 entries and more than 30 different measurements. These measurements cover temperature, how much moisture is in the air, air pressure, details about the wind, how cloudy it is, and how much rain or snow falls.

Data preprocessing was performed in several stages. First, any measurement that was missing for more than 30% of the time was dropped, because we wanted to be sure the data was trustworthy. Remaining missing values were handled using appropriate imputation strategies, taking into account the nature of each variable. Additionally, highly correlated features ($|\rho| > 0.95$) were removed to avoid things being over-emphasized and to help avoid problems with variables affecting each other. For the model specifically designed to predict snowfall, we only used data from December to March. This decision was based on preliminary observations indicating that including non-winter data introduced noise without improving the accuracy of the prediction.

Feature Engineering

Feature engineering was a key component of this study. Rather than relying solely on daily observations, additional features were constructed to capture temporal dynamics and interactions among variables. These included:

- lag features representing values from previous days
- rolling statistics (e.g., 3-day and 7-day averages and standard deviations)
- interaction features such as humidity × cloud cover
- temporal indicators including month, season, and day-of-year

These adjustments helped the model be more aware of how the past affects the present, and the overall build-up of conditions in the atmosphere, both of which are often crucial to understanding rainfall and snowfall.

Model Development and Evaluation

In our study we developed three predictive models, including: a rainfall classification model, a snowfall classification model and a total precipitation regression model. XGBoost was selected as the primary algorithm due to its efficiency and ability to capture nonlinear relationships. The model builds an ensemble of decision trees iteratively, minimizing prediction errors at each step (Chen & Guestrin, 2016). The snowfall model beside XGBoost, in order to align the class imbalance also included Random Forest and LightGBM techniques.

We carefully adjusted the settings of all the forecasting systems using grid search and cross-validation. What we tweaked – the number of 'trees' in the collection, how quickly the system learns, how far each 'tree' can branch, and how to avoid being overly specific to the training data – was all to get the best possible outcome that also works with new data. Figure 1 presents the most influential features for the rainfall prediction model. The results show that humidity and cloud cover are the dominant predictors, particularly through interaction features such as humidity and cloud cover and cloud cover and temperature. This indicates that rainfall is primarily driven by moisture related conditions rather than individual variables in isolation.

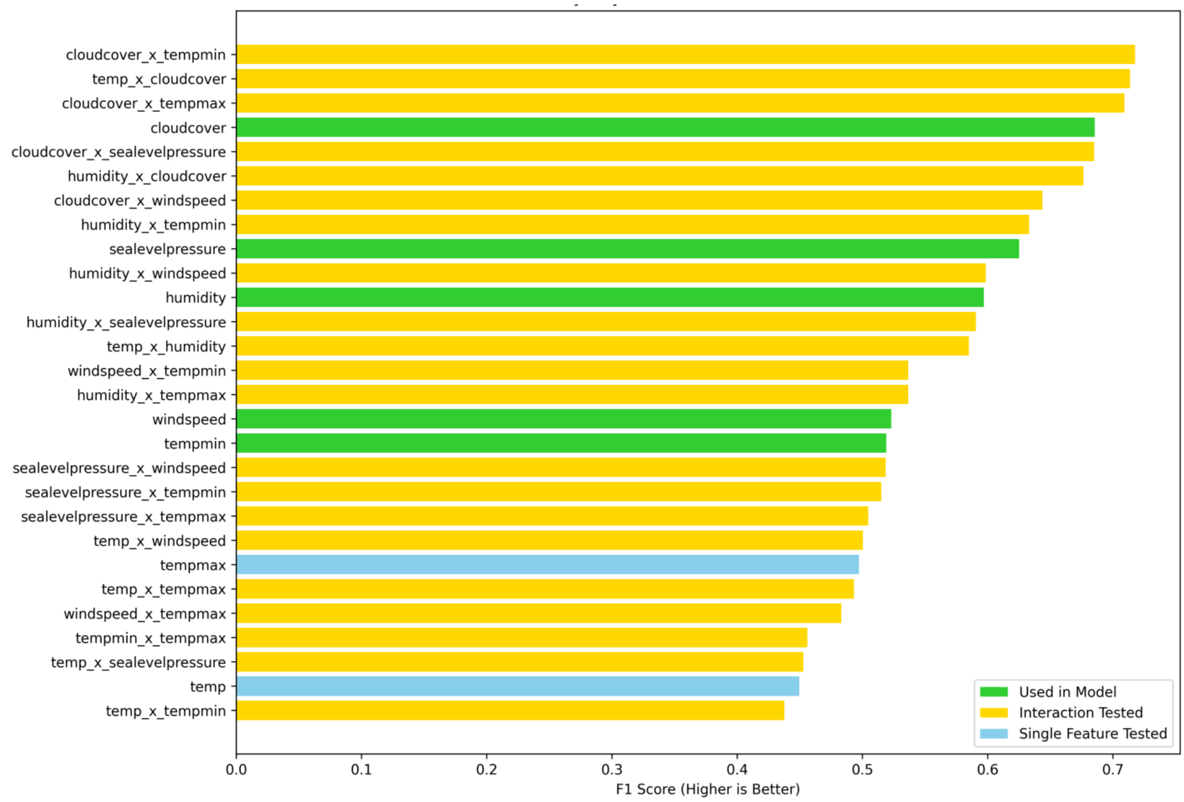

Figure 1. Feature importance and interaction effects for the rainfall prediction model (F1 score).

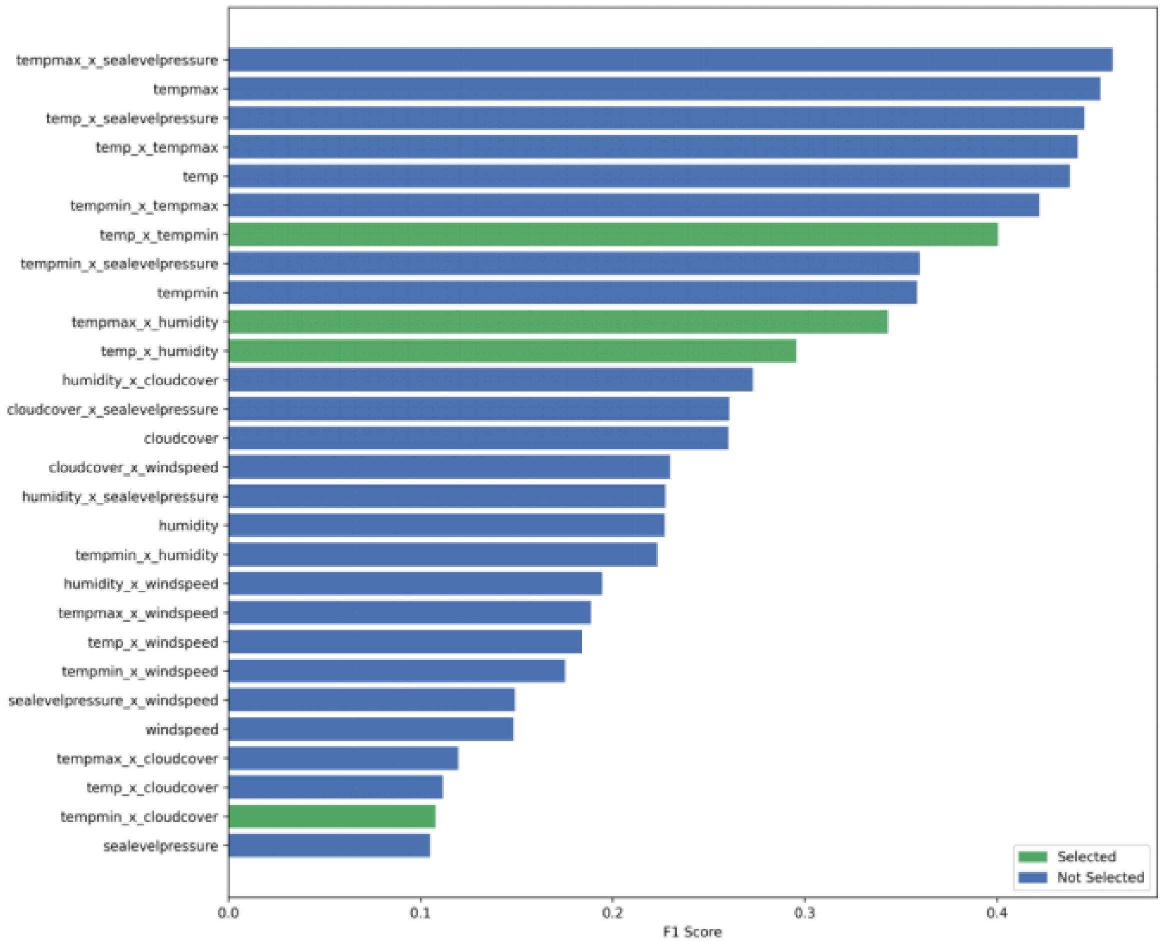

Figure 2. Feature importance and interaction effects for the snowfall prediction model (F1 score).

Compared to snowfall, temperature has a more indirect role, mainly through its interaction with other features. This suggests that rainfall formation depends on combined atmospheric dynamics rather than strict thermal thresholds. As illustrated in Figure 2, the variables that most influence predictions of snowfall are related to temperature. Maximum temperature and sea level pressure interactions are especially significant, a finding in line with what we understand about snowfall's reliance on temperature. The snowfall prediction model's important variables are arranged in a more defined, focused manner than those for rainfall. This is probably because snowfall is strongly tied to the time of year. Interactions between variables are still important but are predominantly those of temperature.

Interestingly, a number of the characteristics chosen for inclusion are not the very best predictors when looked at on their own. This happens because the way we selected features looks at how they work together, rather than how useful each is in isolation. Specifically, some features representing interactions between variables might be fairly average by themselves but greatly improve the model's accuracy when used alongside other characteristics. Furthermore, if strong predictors are very similar to each other, the process may leave out a strong predictor to select a group of characteristics that work better together.

In contrast to the individual models, the complete precipitation model, illustrated in Figure 3, accounts for a wider variety of factors influencing precipitation, and is therefore more broadly applicable, though somewhat less detailed. However, the findings do demonstrate that including interactions and how things change over time makes predictions better and provides a more accurate depiction of what happens in the atmosphere.

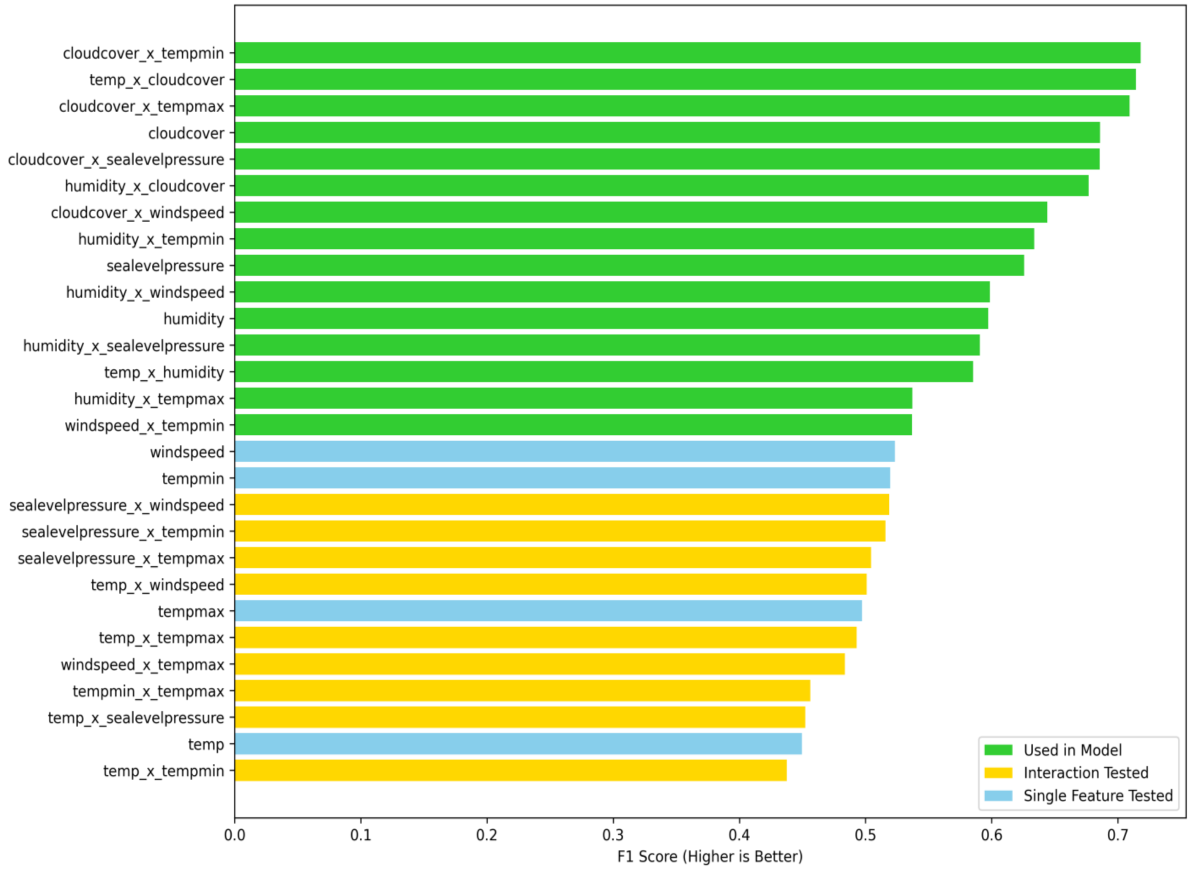

Figure 3. Feature importance and interaction effects across precipitation model (F1 score).

We assessed the models' effectiveness with a five-fold TimeSeriesSplit cross-validation. This method respects the chronological sequence of the data and prevents the unintentional use of future information to predict past outcomes, a crucial consideration in forecasting with time-series.

Results and Discussion

The outcomes demonstrate that all models functioned reliably and at a good standard.

Table 1. Performance of the models

Model	Accuracy
Snowfall	90-97%
Rainfall	81-84%
Percipitation	80-81%

The snowfall model was the most accurate. This is probably because of the more defined seasonality and obvious relationship to temperature. Similar observations have been reported in previous studies (Sanches et al., 2024; Pola & Sekhar, 2021). In the analysis of which features were most important, we identified several primary influences: temperature is the most significant factor in determining snowfall, humidity and cloud cover are tightly connected to rainfall, and previous data points (lag features) greatly improve how well the predictions are made. These findings are in line with what is already known in meteorology and indicate the model is identifying real physical connections, not coincidental associations.

These results support the use of XGBoost for modelling rainfall patterns in complicated climates. In contrast to standard statistical methods, the model is much more easily adjusted to accommodate non-linear relationships and interactions between variables. However, we must also acknowledge certain weaknesses. The effectiveness of the model relies on having good quality, comprehensive data. Furthermore, infrequent or very severe weather occurrences are hard to forecast, as they are not generally found in many historical data sets.

Conclusion

The outcomes of this research validate the efficacy and dependability of XGBoost and ensemble models for predicting rainfall and snowfall over the long term in Skopje, North Macedonia, within a machine learning framework. Accuracy and the ability to apply the predictions to new situations are both notably enhanced when compared to standard methods. In addition to being accurate in their predictions, the results reveal which meteorological variables are most important, and thereby offer a more complete understanding of how precipitation occurs. However, the analysis is still restricted by its reliance on data from only one source and by the absence of precipitation measurements from satellite or radar.

Recommendations

Considering these findings, there are multiple avenues for further investigation. Expanding the data set to include more than just Skopje would be beneficial in determining how applicable the method is to other areas that have different climates. A broader geographical investigation will enhance the solidity of the conclusions. Furthermore, including data from satellites or live observations could enhance the performance of the model, especially in predicting infrequent or severe weather that is not adequately documented in past data. Future research could also assess combined methodologies that link machine learning to physical models. Such a fusion may achieve a stronger equilibrium between predictive precision and how easily the predictions can be explained in terms of physics. And from a functional viewpoint, this system could be transformed into decision making assistance for areas such as farming or city development, both of which would greatly benefit from knowing precipitation levels ahead of time.

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest

Funding

* No funds have been received.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Basic Sciences and Technology (www.icbast.net) held in Konya/Türkiye on April 02-05, 2026.

References

Agarwal, R., & Kumar, V. (2023). Machine learning applications in weather forecasting: A review. *Environmental Modelling & Software, 165*, 105704.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. (2013). Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdisciplinary Reviews: Climate Change, 4*(4), 245–268.

Hoffman, R. N., & Kalnay, E. (1983). Lagged average forecasting: An alternative to Monte Carlo forecasting. *Tellus A, 35*(2), 100–118.

Pola, S., & Sekhar, C. (2021). Prediction of rainfall using extreme gradient boosting (XGBoost). *International Journal of Advanced Research in Engineering and Technology, 12*(3), 1001–1010.

Sanches, R., da Silva, R., & Oliveira, F. (2024). Using XGBoost models for daily rainfall prediction. *Atmospheric Research, 298*, 107200.

Schär, C., Vidale, P. L., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., & Appenzeller, C. (2016). The role of increasing temperature variability in European summer heatwaves. *Nature, 427*(6972), 332–336.

Author(s) Information

Nexhmije Miftari
Mother Teresa University
"1669" nr. 11A, Skopje, North Macedonia
ORCID iD: https://orcid.org/0009-0004-6774-2321

Valon Ahmeti
Mother Teresa University
"1669" nr. 11A, Skopje, North Macedonia
ORCID iD: https://orcid.org/0009-0001-3060-1929

Majlinda Axhiu
Mother Teresa University
"1669" nr. 11A, Skopje, North Macedonia
ORCID iD: https://orcid.org/0000-0002-2567-1623
*Contact e-mail: *majlinda.axhiu@unt.edu.mk*

*Corresponding author(s) contact email

To cite this article:

Miftari, N., Ahmeti, V. & Axhiu, M. (2026). A machine learning framework for long-term precipitation prediction using gradient boosting and ensemble techniques. *The Eurasia Proceedings of Science, Technology, Engineering and Mathematics (EPSTEM), 40,* 391-396.