AI in Climate Science: Leveraging XGBoost and Random Forest for Ocean Temperature Forecasting

Ms. Ruta Prabhu
Department of Information Technology,
Narsee Monjee College of Commerce and Economics,
Mumbai,
Maharashtra,
India

Dr. Anupama Jawale
Department of Information Technology,
Narsee Monjee College of Commerce and Economics,
Mumbai,
Maharashtra,
India

Shivwani Nadar
Department of Information Technology,
Narsee Monjee College of Commerce and Economics,
Mumbai,
Maharashtra,
India

Meenakshi Konar
Department of Information Technology,
Narsee Monjee College of Commerce and Economics,
Mumbai,
Maharashtra,
India

Disha Gandhi
Department of Information Technology,
Narsee Monjee College of Commerce and Economics,
Mumbai,
Maharashtra,
India

Heer Panchal
Department of Information Technology,
Narsee Monjee College of Commerce and Economics,
Mumbai,
Maharashtra,
India

Dr. Ganesh Magar
SNDT Women's University,
Mumbai,
Maharashtra,
India

Abstract:

By examining meteorological and oceanographic variables such as wind patterns, humidity, and air temperature, this study investigates the application of machine learning approaches, Random Forest Regressor and XGBoost Regressor, to forecast sea surface temperature (SST). Using an El Niño dataset, the study applies rigorous preprocessing to integrate spatial-temporal variability and address data discrepancies. Comparative research indicates that XGBoost outperforms Random Forest in terms of prediction accuracy, as seen by higher R2 scores and lower RMSE. The findings provide valuable insights into the ways that oceanic systems are being impacted by climate change and show how advanced machine learning algorithms can capture nonlinear interactions.

Keywords: El Niño, Oceanography, Random Forest, XGBoost, etc

Introduction

To understand comprehensively the effects of climate change in the global distribution of oceanographic characteristics. The oceans generate the climatic system of the earth through heat distribution, carbon dioxide absorption, and weather pattern alteration. It causes the alteration in meteorological parameters such as sea surface temperature (SST), wind pattern, humidity, and air temperature due to growing greenhouse gas emissions, affecting the global weather systems and marine ecosystems [1].

In this study, the relationship between SST and oceanographic and meteorological parameters is established. It uses an El Niño dataset comprising wind, humidity, and air temperature observations [2]. Data reading and preprocessing methods were applied to reconcile differences of data. Its analytical power to handle substantial, high-dimensional data is given by two machine-learning algorithms: Random Forest Regressor and XGBoost Regressor. This paper focuses on to improve the accuracy of SST predictions with advanced machine learning methods and provide insight into impacts of climate change on oceanographic systems [3][4][5][6].

Literature Review

Oceanographic systems are very important for marine ecosystems and the overall control of the world's climate. Sea surface temperature (SST) has an effect on ocean circulation, biodiversity, and the weather. It has always been SST prediction that used traditional statistical methods like linear regression and ARIMA models to uncover trends and patterns associated with phenomena like El Niño and La Niña[7][8]. These methods usually assume linearity, limiting the potential to capture the more complex relationships of oceanographic events.[9]

Innovations in computational technologies have permitted machine learning (ML) techniques to be applied for SST prediction; efficiency of algorithms such as Random Forest (RF) and XGBoost have previously been shown for processing vast datasets and modelling of nonlinear functions. For instance, Random Forest beats the traditional regression models on the accuracy front [10] XGBoost is good in noisy and missing input datasets [11]. These ensemble methods make up the inadequacies of previous statistical methods. While GBT manages large datasets with better accuracy and quite a good speed of calculation [12], RF has the property of managing high-dimension data, avoiding overfitting [13]. However, there is no currently available direct comparison between these methods with respect to SST datasets having spatio-temporal data. One of the complications in SST prediction is the handling of missing data due to irregular measurements or state of the art malfunctions of sensors [14]. Use of advanced preprocessing strategies can reduce such problems.

Methodology

In this research, we used a dataset [15] with 178,080 entries and 12 columns, consisting of temporal and oceanographic information. Important characteristics like "*Zonal Winds*," "*Meridional Winds*," "*Humidity*," "*Air Temp*," and "*Sea Surface Temp*" were first saved as object types, necessitating data cleaning and conversion to numerical formats for precise analysis. Furthermore, specific columns, like "*Humidity*," had data represented by a period (".") denoting missing or invalid values, which were handled using suitable preprocessing methods like imputation or deletion. The "*Year*," "*Month*," and "*Day*" columns were merged to form a complete datetime variable, improving the model's capability to capture temporal patterns.The dataset also includes crucial oceanographic variables, such as "*Latitude*" and "*Longitude*," along with meteorological measures like "*Zonal Winds*" and "*Sea Surface Temperature*," which are vital for understanding the impact of climate change on oceanographic systems. This data preprocessing laid the foundation for the subsequent machine learning analysis and model development.

Data Preprocessing, Feature Selection, and Normalisation

Data preprocessing is essential in transforming unrefined data into a suitable format for training purposes. This procedure includes tidying up the data by dealing with problems like empty values, getting rid of discrepancies, and making sure it is well-organised.

Model Training and Evaluation

To understand comprehensively the effects of climate change in the global distribution of oceanographic characteristics. The oceans generate the climatic system of the earth through heat distribution, carbon dioxide absorption, and weather pattern alteration. It causes the alteration in meteorological parameters such as sea surface temperature (SST), wind pattern, humidity, and air temperature due to growing greenhouse gas emissions, affecting the global weather systems and marine ecosystems [1].

In this study, the relationship between SST and oceanographic and meteorological parameters is established. It uses an El Niño dataset comprising wind, humidity, and air temperature observations [2]. Data reading and preprocessing methods were applied to reconcile differences of data. Its analytical power to handle substantial, high-dimensional data is given by two machine-learning algorithms: Random Forest Regressor and XGBoost

Regressor.Metrics like RMSE and R^2 are used to assess these models, while performance comparisons are displayed through visualisations.The plan is to improve the accuracy of SST predictions with advanced machine learning methods and provide insight into impacts of climate change on oceanographic systems [3][4][5][6].

Comparative Analysis

XGBoost delivers superior performance but needs meticulous tuning and can be costly in terms of computation. Conversely, Random Forest is simpler to adjust and resistant to overfitting but might not consistently reach the level of performance that XGBoost achieves with intricate datasets [18][19][20][21].

Results & Discussion

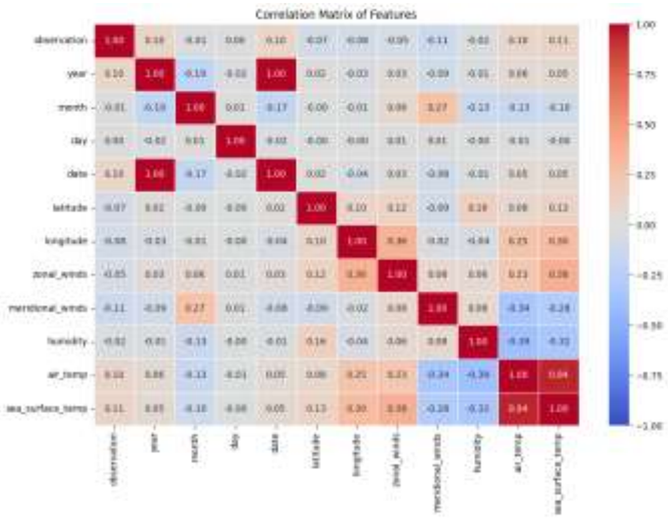


Figure-1: Correlation Matrix of Features

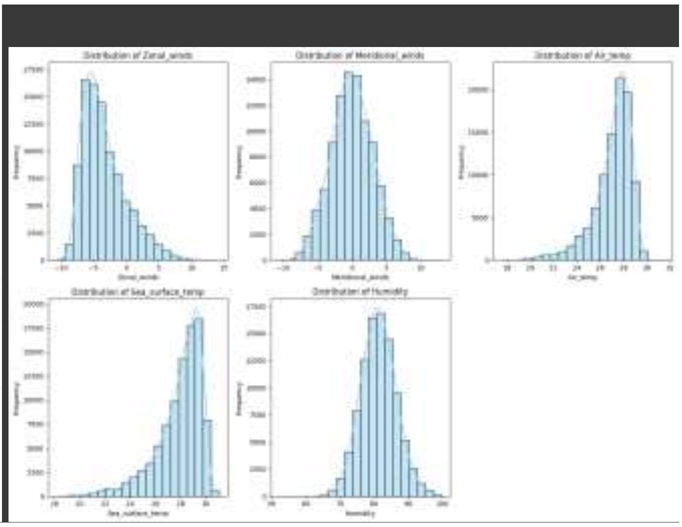


Figure-2: Distribution of Key Features

The above heatmap in ***figure-1*** indicates the relationships among variables in the dataset. For example, *sea_surface_temp* and *air_temp* show strong positive correlation, with a value of 0.94, while *humidity* and *air_temp* feature a strong negative correlation of —0.39. *Latitude* and *zonal_winds* have weak correlations, showing only 0.12, indicating almost no relationship, thereby helping in understanding the patterns as well as leading to feature selection.

The ***figure-2*** thus presents the frequency distributions for *Zonal_winds*, *Meridional_winds*, *Air_temp*, *Sea_surface_temp*, and *Humidity*. Zonal_winds and Meridional_winds are symmetric around zero; while the Air_temp and the Sea_surface_temp are right-skewed, clustering higher values, with Humidity approximating normal variation.

From the ***figure-3***, it can be observed that both models were compared in terms of their Root Mean Squared Error (RMSE). RMSE for XGBoost is lower compared to that of Random Forest, indicating predictive accuracy superior to that of the Random Forest. This simply highlights the performance superiority of XGBoost over random forests in minimising error prediction.

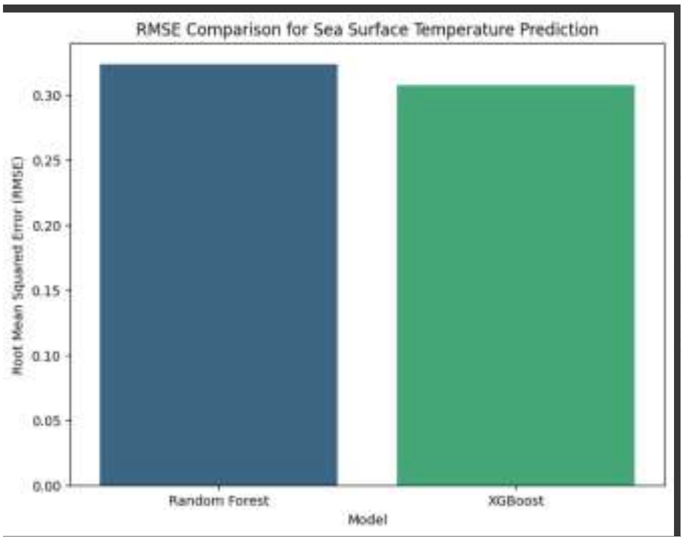


Figure-3 : RMSE Comparison: Random Forest vs XGBoost

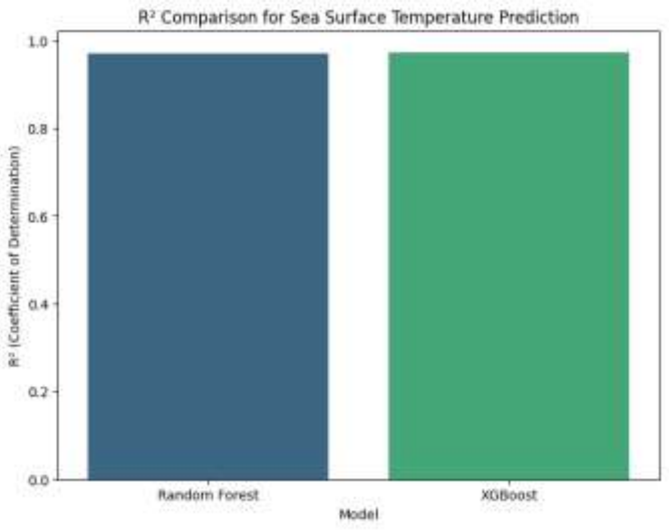


Figure-4: R² Comparison: Random Forest vs XGBoost

Both figures assess model quality according to their R² value, with attention turned towards different ones, as illustrated in ***figure-4***. The two models are being compared according to the R² scores: the proportion of variance explained by the model. Both models yield pretty high R² scores, but the performance with XGBoost is slightly better, which means better predictive accuracy and better fit.

The residuals are presented in ***figure-5*** for the Random Forest model. All these peaks and valleys evaluate model accuracy, clustering mostly around zero, which represents close predictions. However, a much larger spread could also be seen at highly predicted values, meaning the model's performance should still be improved in those particular areas.

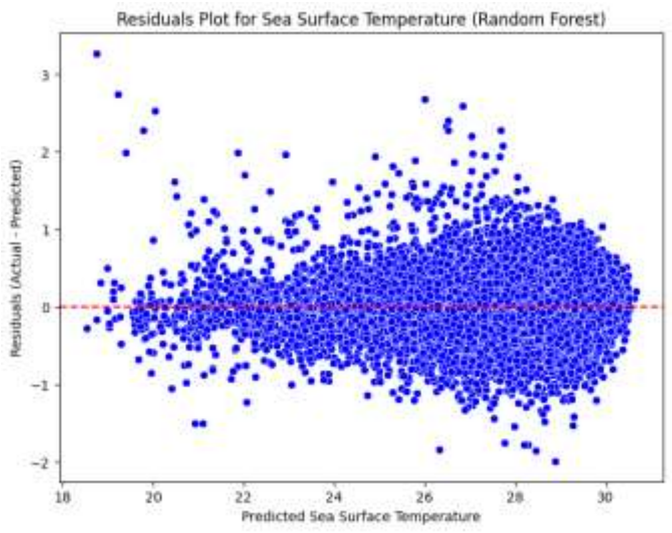
Figure-6 depicts the residuals (differences between predicted and actual values) against predicted values for the XGBoost model. The red dashed line at zero represents the ideal scenario where predictions perfectly match the actual values. The distribution of residuals appears centred around zero, with no significant patterns, indicating that the model is relatively unbiased and performs well across the prediction range. However, some variability in residuals suggests areas where predictions deviate slightly, potentially indicating opportunities for further optimization.

Model Accuracy

Random Forest - RMSE: 0.32,	R²: 0.97
XGBoost - RMSE: 0.31,	R²: 0.98

Conclusion

This research work exposes the dynamic inter-relationship between oceanographic and meteorological variables regulating SST, a key indicator of climate change. It uses a high-quality cleaned dataset from the El Niño events to compare the performances of two machine learning models: Random Forest Regressor and XGBoost Regressor. The performance of the two models is shown to be strong; however, XGBoost was proven to be more precise in prediction with lower RMSE and higher R² than Random Forest models. Residual analysis showed that both the models produced unbiased predictions with residuals clustering around zero, but XGBoost displays better variability in predictions, further underlining its effectiveness.



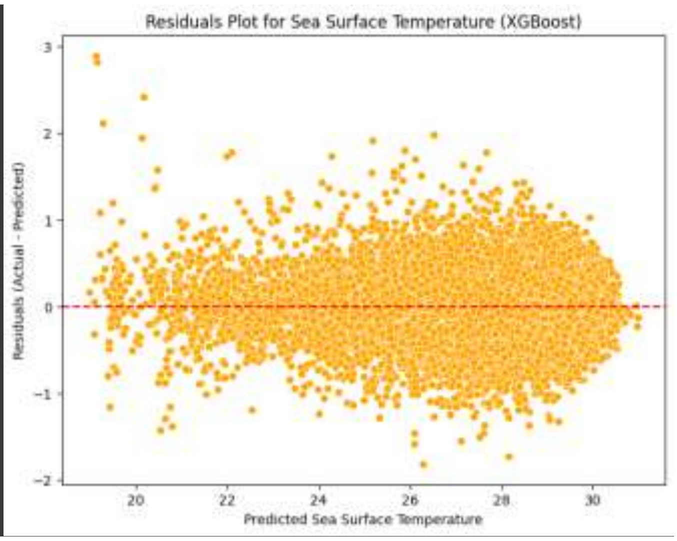


Figure- 5: Residuals Plot for Random Forest Regression Model

Figure 6: Residuals Plot for XGBoost Model

The study highlights the importance of machine learning models in handling the nonlinear complexity of oceanographic datasets. Moreover, it accords more importance to data preprocessing and feature engineering, which cover issues associated with missing data and spatial-temporal variability integration, to improve the model performance. Furthermore, this research fills a gap in the literature by offering complete comparison of ensemble methods and demonstrating advances made in visualisation techniques integrated into predictive frameworks.

Future Scope

These findings, among others, encourage future research into additional oceanographic and meteorological variables for the incorporation of salinity, chlorophyll concentration, and ocean currents to improve the accuracy of the prediction and enable a more holistic understanding of SST dynamics. More detailed spatio-temporal datasets may then be allowed to better refine local predictions, or indeed, through more advanced modelling techniques, such as deep learning and ensemble approaches, further enhance performance. In addition, extending this method in simulating SST for multiple climate change scenarios would be significant in guiding policymakers and environmental planning. Real-time systems established to predict SST and El Niño/La Niña can also make huge contributions to the early warning system and strategies toward climate resilience.

This research forms the basis for taking machine learning in climate science from merely considered propositions to better-informed actions to mitigate climate change impacts on marine environments with enhanced predictions.

References

1. Nagelkerken, I., Allan, B. J. M., Booth, D. J., Donelson, J. M., Edgar, G. J., Ravasi, T., et al. (2023). The effects of climate change on the ecology of fishes. PLOS Climate, 2(8), e0000258.

2. Wang, C. (2019). Three-ocean interactions and climate variability: A review and perspective. \emph{Climate Dynamics}, 53(4), 5119-5136.

3. Henson, S. A., & Sarmiento, J. L. (2011). The role of the ocean in the uptake of anthropogenic carbon dioxide. \emph{Global Biogeochemical Cycles}, 25(3), GB1004.

4. Zhao, M., Held, I. M., & Vecchi, G. A. (2010). Retrospective Forecasts of the Hurricane Season Using a Global Atmospheric Model Assuming Persistence of SST Anomalies. *Monthly Weather Review*, 138(10), 3858-3868

5. Meroni, A. N., Miller, M. D., Tziperman, E., & Pasquero, C. (2017). Nonlinear Energy Transfer among Ocean Internal Waves in the Wake of a Moving Cyclone. *Journal of Physical Oceanography*, 47(8), 1961-1980

6. Ji, Q., Jia, X., Jiang, L., Xie, M., Meng, Z., Wang, Y., & Lin, X. (2024). Contribution of Atmospheric Factors in Predicting Sea Surface Temperature in the East China Sea Using the Random Forest and SA-ConvLSTM Model. *Atmosphere*, 15(6), 670.

7. Chakraborty, K., Choudhury, B., Alom, I., Das, S., Handique, M., & Sharma, N. (2024). Harnessing the Power of LSTM-XGBoost Ensemble Model for Prediction of Sea Surface Temperature Anomalies in the Indian Ocean. In *Data Management, Analytics and Innovation (ICDMAI 2024)* (pp. 501-520). Springer.

8. Ciza Arsène Mushagalusa, Adandé Belarmain Fandohan, & Romain Glélé Kakaï (2024). Predicting species abundance using machine learning approach: a comparative assessment of random forest spatial variants and performance metrics. *Modelling Earth Systems and Environment*, 10, 5145-5171.

9. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937-1967.

10. Renom, M., Capet, X., Maes, C., & Marin, F. (2016). Seasonal variability of the Ekman currents at the entrance of the Gulf of Guinea from PIRATA velocity time series. HAL Open Science

11. Anastassiou, G. A., & Mezei, R. A. (2018). Polynomial Interpolation. In *Numerical Analysis Using Sage* (pp. 117-159). Springer

12. Tal Ezer & Sönke Dangendorf (2022). Spatiotemporal variability of the ocean since 1900: testing a new analysis approach using global sea level reconstruction. *Ocean Dynamics*, 72(1), 79-97.

13. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937-1967.

14. Patel, P., Chaudhary, S., & Parmar, H. (2022). Analyze the Impact of Weather Parameters for Crop Yield Prediction Using Deep Learning. In *Proceedings of the BDA 2022 India* (pp. 249-259). Springer

15. UC Irvine Machine Learning Repository. (n.d.). *El Niño Dataset*. Retrieved from Kaggle

16. Geniesse, C., Chen, J., Xie, T., Shi, G., Yang, Y., Morozov, D., Perciano, T., Mahoney, M. W

17. ., Maciejewski, R., & Weber, G. H. (2024). Visualizing Loss Functions as Topological Landscape Profiles. arXiv preprint arXiv:2411.12136.

18. Tsartsali, E. E., Haarsma, R. J., Athanasiadis, P. J., Bellucci, A., de Vries, H., Drijfhout, S., de Vries, I. E., Putrahasan, D., Roberts, M. J., Sanchez-Gomez, E., & Roberts, C. D. (2022). Impact of

resolution on the atmosphere–ocean coupling along the Gulf Stream in global high resolution models. Climate Dynamics, 58(3), 3317-3333.

19. Hoegh-Guldberg, O., & Poloczanska, E. S. (2017). Editorial: The Effect of Climate Change across Ocean Regions. Frontiers in Marine Science, 4, 361.

20. Gattuso, J.-P., Magnan, A. K., Bopp, L., Cheung, W. W. L., Duarte, C. M., Hinkel, J., ... & Williamson, P. (2018). Ocean Solutions to Address Climate Change and Its Effects on Marine Ecosystems. Frontiers in Marine Science, 5, 337.

21. Poloczanska, E. S., Brown, C. J., Sydeman, W. J., Kiessling, W., Schoeman, D. S., Moore, P. J., ... & Richardson, A. J. (2016). Responses of Marine Organisms to Climate Change across Oceans. Frontiers in Marine Science, 3, 62.

22. Marbà, N., & Duarte, C. M. (2017). Effects of climate change across ocean regions. Frontiers in Marine Science, 4, 1-11.