Predicting Ocean Current Temperature Off the East Coast of America with XGBoost and Random Forest Algorithms Using Rstudio

Lulut Alfaris[1]*, Anas Noor Firdaus[1], Ukta Indra Nyuswantoro[2], Ruben Cornelius Siagian[3], Aldi Cahya Muhammad[4], Rohana Hassan[5], Rodulfo T. Aunzo, Jr.[6], Reza Ariefka[7]

[1]*Department of Marine Technology, Politeknik Kelautan dan Perikanan Pangandaran*
Babakan, Kec. Pangandaran, Kab. Pangandaran, Jawa Barat 46396 Indonesia
[2]*Department of Structural Engineering, Asiatek Energi Mitratama*
Pakuwon Tower Unit 10F, Tebet, South Jakarta City, Jakarta 12870 Indonesia
[3]*Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan*
Jl. Williem Iskandar Psr. V, Medan Estate, Kabupaten Deli Serdang, Sumatera Utara 20221 Indonesia
[4]*Department of Electrical and Electronic Engineering, Islamic University of Bangladesh*
4th Floor, Dr.M.A Wazed Miah building, Islamic University, Bangladesh
[5]*Institute for Infrastructure Engineering and Sustainable Management (IIESM), Universiti Teknologi MARA,*
40450 Shah Alam, Selangor, MalaysiaShah Alam, Selangor, Malaysia
[6]*Assistant Professor IV, Visayas State University Isabel*
Isabel, Leyte, Philippines
[7]*Research Institute STKIP Muhammadiyah OKU Timur*
Jl. Pujorahayu, Harjo Winangun, Ogan Komering Ulu Timur, Sumatera Selatan 32382 Indonesia
Email: lulut.alfaris@kkp.go.id

Abstract

This research investigates the comparative predictive efficacy of two leading machine learning methodologies, specifically the XGBoost and Random Forest models, in estimating ocean temperature dynamics in the TS Gulf Stream and Labrador Current regions along the east coast of North America. Using annual temperature datasets and relevant oceanographic parameters, the data is carefully processed, cleaned and sorted into training and test subsets via the RStudio Platform. The performance evaluation model is carried out using predetermined machine learning assessment criteria, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and R-squared. The results show the superiority of the XGBoost model compared to Random Forest in terms of prediction accuracy and minimizing prediction errors. The XGBoost model shows lower MSE values and higher R-squared values than the Random Forest model, indicating its better capacity in explaining data variations. XGBoost consistently provides more accurate predictions and shows higher sensitivity in identifying important factors influencing ocean temperature fluctuations than Random Forest. This research significantly improves understanding and prognostic capabilities regarding ocean temperature dynamics in the TS Gulf Stream and Labrador Current regions. Empirical evidence underlines the efficacy of the XGBoost model in predicting ocean temperatures in the studied region. Continuous model evaluation and parameter refinement for both methodologies is critical to establishing standards for optimal prediction performance. The findings of this research have implications for the fields of oceanography and climate science, and offer potential pathways to comprehensively understand and mitigate the impacts of climate change on marine ecosystems.

Keywords: *Forecasting, Machine learning methods, Model performance metrics, Predictive accuracy*

Introduction

The ocean temperature phenomenon in the TS Gulf Stream and Labrador Current off the east coast of North America is an ecological event that has significant implications for marine ecology (Chen, 2022). The Gulf Stream is a warm and swift ocean current that originates in the Gulf of Mexico and flows along the east coast of North America before turning towards Europe (Williams, 2021). The Labrador Current, on the other hand, is a cold current that flows southward from the Arctic Ocean along the coast of Labrador and Newfoundland before merging with the Gulf Stream (Lochte *et al*., 2019). The interaction between these two ocean currents creates a dynamic ocean environment with significant temperature variations that impact marine life *(*Jacobs *et al*., 2021). The warm waters of the Gulf Stream provide an ideal habitat for species such as sea turtles, dolphins, and sharks, while the cold waters of the Labrador Current support species such as cod, haddock, and capelin (Pannozzo, 2023).

*) Corresponding author

https://ejournal.undip.ac.id/index.php/ijms
DOI: 10.14710/ik.ijms.29.2.273-284
Received : 28-01-2024
Accepted : 05-04-2024

Understanding the dynamics of ocean currents and their impact on temperature is critical for predicting changes in marine ecosystems and developing accurate predictive models (Capotondi *et al*., 2019). Accurate predictive models can help scientists and policymakers develop effective conservation strategies and mitigate the impacts of climate change on marine life (Urban, 2019). The XGBoost algorithm is one such model that can handle large and complex data sets and is an effective tool for predicting ocean temperatures. The algorithm uses gradient boosting to construct a sequence of regression trees that iteratively add new trees to the model while minimizing a loss function (Devos *et al*., 2020). The amalgamation of predictions from all individual trees culminates in the final prediction, a crucial step that guarantees the accuracy of the XGBoost model. The development of precise predictive models, such as XGBoost, holds paramount importance in comprehending the intricate dynamics of ocean temperature and its repercussions on marine life (Wolff *et al*., 2020). These models serve as invaluable tools for researchers and policymakers, enabling them to formulate effective conservation strategies and address the impacts of climate change on marine ecosystems. Ultimately, these efforts contribute to ensuring the sustainability of our oceans for the well-being of future generations.

Utilizing the XGBoost algorithm, a widely adopted ensemble method in machine learning that utilizes gradient boosting to construct sequences of regression trees, this study aimed to predict ocean current temperatures at two specific locations off the east coast of North America—the Gulf Stream and Labrador Current. The selection of XGBoost was deliberate, chosen for its prowess in managing large and intricate datasets. When coupled with Gradient Boosting and the xgboost library, the XGBoost algorithm exhibited remarkable accuracy in forecasting ocean current temperatures, evident in the lower Root Mean Square Error (RMSE) *(*Duan *et al*., 2023).

Implementing the Gradient Boosting method on ocean current temperature data at specified locations, the study divided the dataset into training and testing sets with a 70:30 ratio. The model was developed using the xgboost package, and predictions were made on the testing data (Qiu *et al*., 2022). The Support Vector Machine (SVM) algorithm emerged as an effective method for predicting ocean current temperatures within the given data (Khan *et al*., 2021).

Representing two significant ocean currents along the eastern shores of North America are the Gulf Stream and the Labrador Current (Gonçalves Neto *et al*., 2023). Originating in the warm waters of the Gulf of Mexico, the Gulf Stream flows northeastward along the eastern coast of the United States before turning eastward toward Europe (Bruera *et al*., 2023). Renowned as one of the world's strongest ocean currents, it is characterized by warm, clear, blue water and is situated approximately between 35°N to 45°N latitude and 75°W to 60°W longitude. The Labrador Current is a cold ocean current that travels southward along the east coast of Canada, originating in the frigid waters of the Labrador Sea (Board, 2021). These two currents play pivotal roles in shaping the marine ecosystems and climates of the region, with the Gulf Stream's warm, energetic flow and the Labrador Current's cold influence contributing to the distinct characteristics of the coastal waters (Trossman and Palter, 2021).

The primary objective of the study was to utilize the XGBoost algorithm in machine learning to predict ocean current temperatures at two specific locations off the east coast of North America: the Gulf Stream and Labrador Current. Its focus was to demonstrate the precision and efficiency of the XGBoost algorithm in handling intricate and expansive datasets. This study offers several benefits. Firstly, it successfully showcases the XGBoost algorithm's ability to accurately forecast ocean current temperatures at precise locations. Secondly, it emphasizes the crucial role of selecting the appropriate machine learning method in managing complex datasets and enhancing accuracy. Thirdly, it provides a practical framework for implementing the XGBoost algorithm in R programming through the xgboost package. Lastly, the study explores alternative algorithms such as Random Forest and SVM for forecasting ocean current temperatures, offering a comparative analysis of their effectiveness in handling the given data. This research substantiates the utility and effectiveness of the XGBoost algorithm in predicting ocean current temperatures, highlighting the importance of choosing the right machine learning approach for intricate datasets. Its findings have broad applicability across various domains, particularly in fields like climate research and oceanography, where precise forecasts of ocean currents have significant implications.

Materials and Methods

The xgboost algorithm model

In this research, an in-depth exploration of the xgboost model for predicting ocean temperature has been carried out, employing a training dataset $D = (x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$. The xgboost model, inspired by the methodology proposed by (Wang and Guo, 2020), meticulously constructs a sequence of regression decision trees $f_m(x)$. This construction is achieved by minimizing the loss function:

$$L^{(m)} = \sum_{i=1}^{n} l\{y_i, \hat{y}^{(m-1)} + f_m(x_i)\} + \Omega(f_m) \quad (1)$$

Integral to the model's efficacy is the regularization rule $\Omega(f_m)$, a critical mechanism controlling the complexity of the decision tree and guarding against overfitting—a technique commonly employed in xgboost as highlighted by (He *et al.*, 2021). This regularization term optimizes the linear regression loss function $l(.,.)$ when applied to the decision tree f_m. In each iteration (m), the predicted values $\hat{y}^{(m-1)}$ from the preceding iteration $(m-1)$ guide the training of the decision tree f_m. The regularization term $\Omega(f_m)$ serves as a deterrent to overfitting by penalizing the tree's complexity. Among the commonly adopted regularization rules in xgboost is:

$$\Omega(f_m) = \Upsilon.T + \frac{1}{2}.\lambda \sum_{j=1}^{T} w_j^2 \quad (2)$$

Where (Υ) and (λ) stand as predetermined hyperparameters, T denotes the number of leaves in the tree (f_m), and (w_j) represents the weight of the j-th leaf. In each iteration, a new regression tree $f_m(x)$ is introduced by minimizing the loss function $L^{(m)}$ through the gradient descent algorithm. The culmination of the xgboost model's predictive power on a feature vector x is articulated by $\hat{y}(x) = \sum_{m=1}^{M} f_m(x)$, where M signifies the total number of regression trees meticulously crafted by the xgboost model.

The computational realization of this algorithm in R involves a meticulous process. Hyperparameters (Υ), (λ) and M are meticulously initialized. Subsequently, an array is created to capture the predictions of each regression tree for every sample. Each iteration, spanning from 1 to M, involves the computation of gradients and hessian for each sample using the linear regression loss function. This is followed by the meticulous construction of a new regression tree, adhering to regulation rules that encompass data separation based on features and the calculation of weights for each leaf. The resulting predictions for each sample are saved for subsequent iterations. The xgboost model prediction for the feature vector x is calculated, ultimately providing the output $\hat{y}(x)$.

Random forest algorithm model

Random forest algorithm mathematical model

The Random Forest algorithm, a versatile machine learning tool employed for classification, regression, and diverse tasks, was investigated in this study (Shanmugasundar *et al*., 2021). Operating as an ensemble learning method, the algorithm constructs numerous decision trees and amalgamates their predictions (Reddy *et al*., 2020). Mathematically, the Random Forest model can be expressed as follows:

Consider $T_1, T_2, \ldots T_n$ as n decision trees, with each tree (T_i) crafted on a random subset of the training data. At every node in the tree, the data is partitioned based on the optimal feature. The output of the Random Forest model is given by the equation:

$$\hat{y} = \arg\max_y \sum_{i=1}^{n} w_i \cdot [y = T_i(x)] \quad (3)$$

Where $\hat{y}$ signifies the predicted class or value, y represents the set of potential classes or values, x denotes the input data, w_i stands for the weight of tree i, and $y = T_i(x)$ is an indicator function that yields 1 if the tree T_i predicts class or value y for input x, and 0 otherwise.

The weights w_i are determined by the accuracy of each tree, ensuring that better-performing trees are assigned higher weights. This strategic weighting diminishes the influence of weaker trees, ultimately enhancing the overall accuracy of the Random Forest model (Rong *et al*., 2020).

Algorithm the mathematical model of the Random forest algorithm in the R program

The study began by collecting historical temperature data, which was then divided into two groups: training data and test data. The training data serves to train the model, while the test data is used to test how well the model works (Shokri *et al*., 2017). Each decision tree is generated through a process of random subsampling of the training data and random selection of variables (features) considered at each split node (Liu *et al*., 2020).

The splitting of the training data is done by considering the threshold value of the selected variables, and the goal is to maximize the diversity in the predicted values at each branch (node) and minimize the diversity in the predicted values at each leaf under that branch. The splitting process continues until the number of nodes reaches the maximum limit or after the stopping criterion is met, *i.e.* when there are no more significant splits (Ishwaran, 2015).

After all the decision trees are generated, temperature estimation is performed by means of each decision tree generating temperature predictions for each test data. The temperature predictions from each decision tree are then averaged to produce the final temperature prediction (Birant, 2011). The prediction error is measured using metrics such as MSE *(mean squared error)* or MAE *(mean absolute error)* (Hodson, 2022). Some model parameters, such as the number of decision trees, the number of features considered at each splitting node, and the stopping criteria, can be adjusted to improve model performance (Song and Ying, 2015).

Random forest vs XGBoost Algorithm in a box plot to evaluate the performance of the algorithmic model

This section provides an interpretation of the summary statistics derived from the actual data, specifically focusing on the maximum temperature and minimum temperature recorded at specific point locations within the Gulf Stream and Labrador Current along the east coast of North America. Table 1 presents monthly temperature statistics for the maximum temperature, focusing on point locations in the Gulf Stream and Labrador Current along the east coast of North America. The data includes minimum and maximum values, quartiles, median, and mean temperatures. For instance, the median maximum temperature observed is 6.50, and the corresponding median minimum temperature is 24.13, indicating the central tendencies of the temperature distribution in the specified locations. The table provides insights into the variability and range of temperatures with minimum and maximum values across the recorded months.

Table 2 shows monthly temperature statistics for the minimum temperature at point locations in the Gulf Stream and Labrador Current along the east coast of North America. The data includes key summary measures such as minimum and maximum values, quartiles, median, and mean temperatures. For instance, the median minimum temperature recorded is 6.50, corresponding to 22.97, providing insights into the central tendencies and variability in the distribution of minimum temperatures across the observed months. The table captures the range and distribution characteristics, emphasizing the monthly variations in minimum temperatures at the specified locations.

The plot Figure 1 shown is the monthly seasonal pattern of maximum and minimum temperatures at the study site. The plot uses pre-processed data, where the temperature data has been aggregated by month to obtain the average maximum and minimum temperatures. The plot displays an x-axis representing the months (from 1 to 12) and a y-axis representing the average temperature. There are two types of data displayed in the plot, namely maximum temperature and minimum temperature. Each type of data is displayed with a different color on the bar-chart. The plot shows the pattern of maximum and minimum temperature changes throughout the year. The boundary line between the two types of data shows the difference between the maximum and minimum temperatures in each month. This plot provides a visual representation of the seasonal changes in temperature, where it can be seen whether the temperature is higher or lower in certain months.

Table 1. Monthly Temperature Statistics in maximum temperature

Month	Temperature Max
Min. : 1.00	Min. : 21.27
1st Qu. : 3.75	1st Qu. : 22.17
Median : 6.50	Median : 24.13
Mean : 6.50	Mean : 24.62
3rd Qu. : 9.25	3rd Qu. : 27.04
Max. : 12.00	Max. : 28.84

Table 2. Monthly Temperature Statistics in minimum temperature

Month	Temperature Min
Min. : 1.00	Min. : 19.22
1st Qu. : 3.75	1st Qu. : 20.64
Median : 6.50	Median : 22.97
Mean : 6.50	Mean : 23.40
3rd Qu. : 9.25	3rd Qu. : 26.30
Max. : 12.00	Max. : 28.14

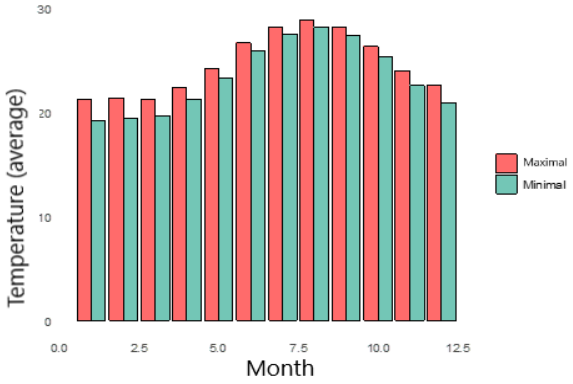


Figure 1. Seasonal month pattern of maximum and minimum temperature

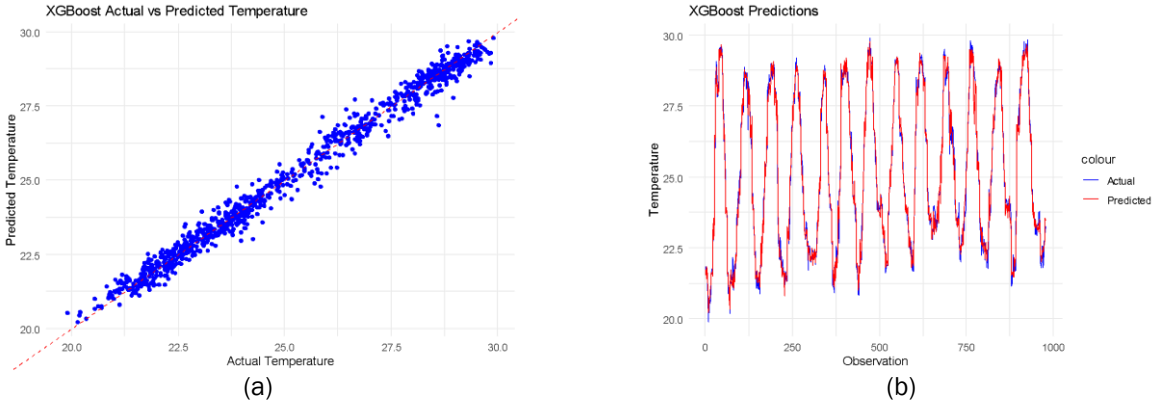


Figure 2. Prediction of ocean current temperature using gradient boosting algorithm

Results and Discussion

In this section, we present a comprehensive report detailing the application of the xgboost algorithm in conjunction with the R programming language to forecast ocean current temperatures at two distinct locations – the Gulf Stream and Labrador Current situated off the east coast of North America. The dataset utilized in this study originates from NASA's power access datasets, renowned for furnishing extensive and reliable information essential for precise temperature predictions in ocean currents. Employed as a machine learning technique, the xgboost algorithm is adept at forecasting intricate and voluminous datasets (Asselman *et al*., 2023). This research contributes valuable insights into the accuracy of predicting ocean current temperatures at both the Gulf Stream and Labrador Current locations through the utilization of the xgboost algorithm. The ensuing output generated by this algorithm is elucidated as follows.

The presented research showcases the remarkable predictive capabilities of the XGBoost model through a meticulously analyzed dataset. In Figure 2.a, a compelling visual representation is provided, juxtaposing the actual temperature (Actual) with the XGBoost model's predicted temperature (Predicted). The alignment of each blue dot, corresponding to a pair of actual and predicted temperatures, vividly illustrates the model's ability to closely replicate true values. The proximity of these dots to the red dashed line serves as a visual indicator of prediction accuracy, with closer dots denoting more precise predictions. This reference line establishes the ideal relationship between actual and predicted temperatures, serving as a valuable tool for assessing the model's precision.

The close alignment of the blue dots to the reference line strongly validates the accuracy of the XGBoost model, affirming its ability to make precise predictions of temperature. Conversely, scattered or distant dots in the plot suggest a higher error rate in the model's predictions.

In Figure 2.b, the sequence of test data on the x-axis is juxtaposed with the temperature scale (encompassing both actual and predicted values) on the y-axis. The blue line represents actual temperatures, while the red line portrays predicted values. A close alignment between these lines

signifies the model's adeptness at accurate predictions, emphasizing its reliability. Conversely, a noticeable difference between the two lines indicates a significant error in temperature prediction.

This graphical representation serves as a comprehensive tool to evaluate the XGBoost model's performance in predicting temperatures. It not only facilitates a thorough understanding of temperature variations within the dataset but also provides valuable insights into the model's predictive capabilities. Assessing the closeness between the red and blue lines aids in determining the model's accuracy, making this plot an essential visual aid for comprehending the model's predictive efficiency.

In this research, an in-depth analysis of the attributes of the XGBoost model (Table 3.) was conducted. The use of external marker handles focused on managing and manipulating the XGBoost booster object externally. The raw data related to the XGBoost model has an attribute length of 385,622. The model training process is performed with only one iteration, called niter, indicating that the iteration is sufficient. The model training evaluation log is stored in a data.table object named evaluation_log. The XGBoost model invocation information, or call, is used as a reference for model parameters and arguments. XGBoost model training parameters are represented by a list object called params. List callbacks include additional functions used during XGBoost model training. There are 5 features used in the model, represented by feature_names. The number of features (nfeatures) in the model is 1, indicating the use of only one feature.

In the statistical analysis of the prediction model, the following temperature values were observed: The observed temperature minimum was 19.90, while the predicted minimum was slightly higher at 20.22. At the 25th percentile, both actual and predicted temperatures reached about 22.83 and 22.86, respectively, representing the lower half of the range. At the 50th or median percentile, the actual and predicted temperatures converge close to 24.65 and 24.66, respectively, indicating the average temperature value. When advancing to the 75th percentile, the actual and predicted temperatures increase to about 25.19 and 25.18, respectively. At the upper end, the maximum observed temperature reaches 27.90, with the predicted temperature at a slightly lower value of 27.85, while the highest values recorded are 29.90 and 29.79 for the observed and predicted temperatures, respectively.

The program used to implement Random Forest algorithm in predicting time series values on ocean current temperature data off the east coast of North America. The prediction results are displayed in the form of a plot with the original data. In this study, the graph illustrates the relationship between the observed and predicted values using the Random Forest model.

In Figure 3, the close relationship between the plot and the dashed red diagonal line indicates a strong correlation between the predicted and observed values. This indicates a linear relationship between the input variables (Year, Month, Date, Maximum Temperature, Minimum Temperature) and the output variable (Temperature). A closer fit of the data to the diagonal line indicates higher prediction accuracy in the temperature model. Variations in the distribution of the data indicate diverse predicted values for each observation. An even distribution around the diagonal line indicates a strong predictive ability in handling temperature fluctuations by the model. Deviations above or below the line indicate the difference between the predicted and observed values-positive deviations indicate overestimation, while negative deviations indicate underestimation by the model. A wider spread may indicate unexplained variability in the model inputs.

The fit between the dashed red line (Predicted) and the blue line (Actual) emphasizes the expertise of the Random Forest model in temperature prediction. Closer proximity between these lines indicates higher prediction accuracy, while differences or non-uniformities indicate mismatches influenced by unaccounted factors such as extreme weather conditions. Comparison of actual and predicted temperatures at statistical points-minimum, quartile, median, mean and maximum-reveals the consistency between model predictions and actual temperatures across a wide range. These collective findings emphasize the model's precision in temperature prediction. Evaluation of the model's overall performance involves assessing the summary statistics obtained through the command "summary(rf_model)," which reveals the accuracy and error rate in temperature prediction.

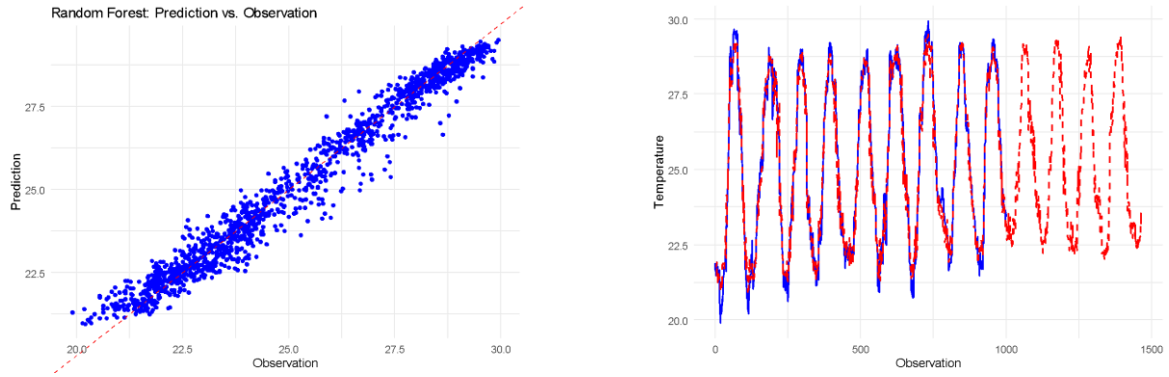
The model evaluation (Table 4.) results show that the Random Forest model has an error rate (mse) of 100 and a fit rate (rsq) of 100. MSE (Mean Squared Error) measures how well the model predicts temperature, with lower values indicating more accurate predictions. R-squared (R^2) describes how well temperature variations can be explained by the model, with higher values indicating a better model in explaining temperature variations. There are 5 variables that are considered important in predicting temperature based on the results of the Random Forest model. These important variables can provide insight into the factors that influence temperature. There is 1 tree (ntree) used in this model. The number of variables used in each split node selection is 1 (mtry).

Table 3. XGBoost Model Attributes

Attribute	Description
Handle	External marker for the XGBoost booster object. Used for managing and manipulating the model.
Raw	Raw data associated with the XGBoost model. The length of this attribute is 385,622.
Niter	The number of iterations performed during model training. Its value is 1, indicating only one iteration.
Evaluation_log	Evaluation log related to the model training. It is a data.table object containing evaluation information.
Call	Information about the XGBoost model call. Used as a reference for model parameters and arguments.
Params	Parameters used in the XGBoost model training. A list object with parameter values.
Callbacks	List of callbacks used during the XGBoost model training. Functions that perform additional actions.
Feature_names	Names of the features used in the model. There are 5 features used.
Nfeatures	The number of features used in the model. Its value is 1, indicating one feature is used.

Table 4. Evaluation and Model Details for Temperature Prediction in random forest model

Details	Value
Total data	3419
Predicted Variables	Temperature
Model Evaluation	
- *MSE (Mean Squared Error)*	100
- *R-squared (R^2)*	100
Significant Variables	5
Number of Trees and Number of Variables Used	
- *Number of Trees (ntree)*	1
- *Number of Variables (mtry)*	1



(a) (b)

Figure 3. Random forest prediction for Gulf Stream and Labrador Current Temperature

For this study, the data used was extracted from an Excel file and organized into a data frame using the data.frame () function. To split the data into training and testing sets with a 70:30 ratio, the create Data Partition () function from the caret library was utilized. Two models were then trained using the training data: Gradient Boosting with xgboost() and Random Forest with train() from the caret library. The testing data was used to make predictions with both models, and the RMSE values and Confidence Intervals (CI) were calculated for each model.

In Figure 4, each plotted point represents a pair of actual and predicted temperature values, which are identified by different types according to the XGBoost and Random Forest models, distinguished by the color of the point. In this plot there is a dashed line (geom_abline) depicting a diagonal line with a slope of 1 and an intercept of 0, illustrating an ideal scenario where the predicted values are parallel to the actual values. The closeness of these points to the line reflects the accuracy of the predictions to the actual values. This visual aid aims to assess the precision of the XGBoost and Random Forest regression models in predicting temperature values. Points that are closer to the diagonal line indicate more precise and accurate predictions, while scattered points indicate lower prediction accuracy. This plot analysis facilitates the evaluation and comparison of the performance of XGBoost and Random Forest models in temperature prediction.

Table 5. Performance Comparison of XGBoost and Random Forest Regression Models for Temperature Prediction

Model	MSE	R-squared
XGBoost	0.228929	0.992583
Random Forest	0.601143	0.95681

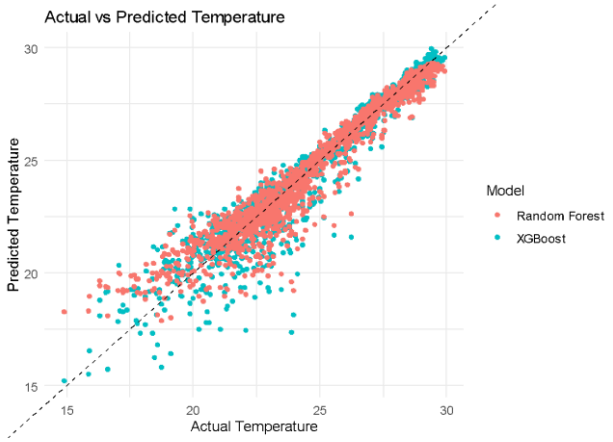


Figure 4. Predicted temperature vs actual temperature: Random Forest vs XGBoost Algoritm

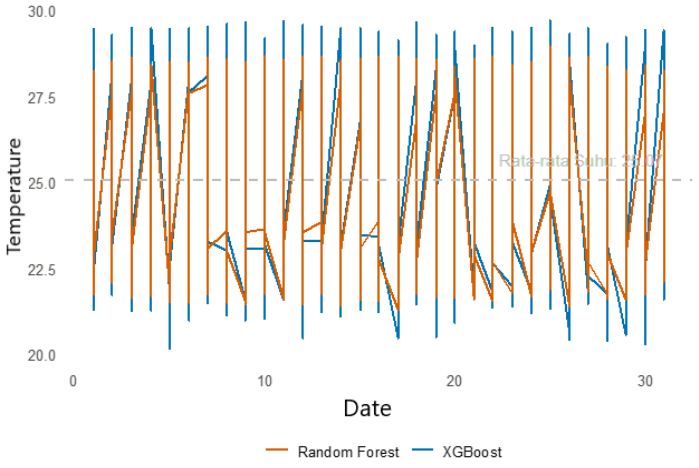


Figure 5. Random forest vs XGBoost Algoritm in line graph

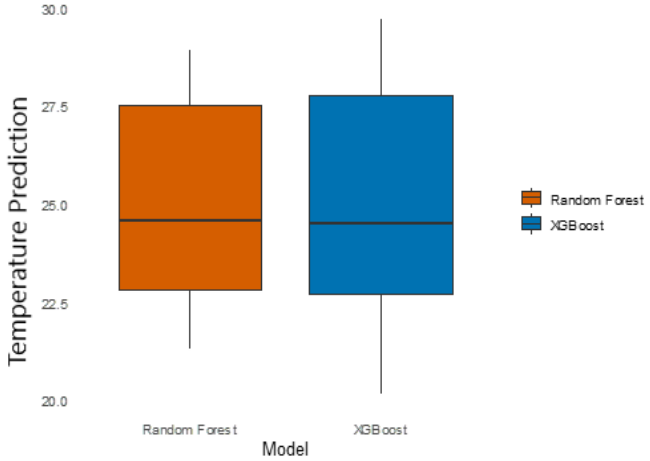


Figure 6. Random forest vs XGBoost Algorithm in a box plot to evaluate the performance of the algorithmic model

In the results of this study, Figure 5 shows the comparison between two prediction models, namely XGBoost and Random Forest, in predicting temperature based on date. On the x-axis (horizontal) is the date, while on the y-axis (vertical) is the temperature value. In this graph, there are two lines representing the predictions of each model. The blue line represents the prediction using the XGBoost model, while the red line represents the prediction using the Random Forest model. This graph helps in comparing the performance of the two models in predicting the temperature. If the prediction line is close to the true value (observed temperature value), it indicates that the model provides an accurate prediction. If there is a difference between the prediction line and the actual value, it indicates the level of prediction error of each model.

In Figure 6, the resulting plot is a box plot comparing temperature predictions using two models, XGBoost and Random Forest. The x-axis displays the two model categories: XGBoost and Random Forest, while the y-axis shows the temperature prediction values. Each box in the plot represents the distribution of temperature predictions for each model, with the horizontal line inside indicating the median of that distribution. Whiskers, represented by vertical lines above and below the boxes, indicate the range of the data (minimum and maximum values). The presence of points outside the whiskers indicates the presence of outliers (extreme values) in the temperature predictions. This plot aims to compare the temperature prediction distribution between the XGBoost and Random Forest models. Similar positions of the boxes and aligned medians indicate similar predictions from both models. Significant differences in the boxes or medians indicate variations in temperature predictions between the two models.

The evaluation results (Table 5.) show that the XGBoost model has a lower mean squared error (MSE) of 0.2289287 compared to the Random Forest model (MSE = 0.6011426). The lower MSE indicates that the XGBoost model has a smaller prediction error rate in estimating temperature. The R-squared value generated by the XGBoost model (0.9925834) is also higher than the Random Forest model (0.9568096). The higher R-squared indicates that the XGBoost model can better explain the variation in the observed data compared to the Random Forest model. In the context of this study, the evaluation results show that the XGBoost model provides better performance in predicting temperature compared to the Random Forest model.

This study delves into the comparison of two prominent machine learning methods, XGBoost and Random Forest, to assess their effectiveness in predicting variations in ocean temperature along the east coast of North America, specifically in the TS Gulf Stream and Labrador Current regions (Ferchichi *et al*., 2022). Utilizing RStudio, the researchers employed a comprehensive dataset consisting of annual temperature records from these regions (Moraga and Baker, 2022). In the XGBoost approach, various oceanic parameters such as air temperature, atmospheric pressure, wind characteristics, and other relevant factors were utilized as input variables (Cui *et al*., 2024). The dataset underwent preprocessing and segmentation into training and testing subsets, enabling the XGBoost model to be trained on oceanic parameters to predict sea temperature. Evaluation metrics such as accuracy, precision, recall, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were employed to assess the model's performance (Naser and Alavi, 2020).

The results indicate that the XGBoost model, developed using RStudio, demonstrated accurate predictions with minimal error, showcasing its proficiency in identifying patterns within annual ocean temperature data in the specified regions (Hussain *et al*., 2021). The study incorporated the Random Forest method, using annual temperature data as model input and following a similar segmentation process (Wei *et al*., 2020). The Random Forest model exhibited commendable performance in predicting ocean temperature variations based on the provided annual temperature data (Azari *et al*., 2022). The analysis of key features within the Random Forest model further enhanced the understanding of factors influencing sea temperature in the studied regions (Oukawa *et al*., 2022). The research effectively compared the predictive performance of XGBoost and Random Forest methods in forecasting sea temperature, offering valuable insights for advancing understanding and predictive capabilities regarding ocean temperature changes in the TS Gulf Stream and Labrador Current regions.

Despite the apparent superiority of the XGBoost model in this context, the study emphasizes the need for further assessments and parameter adjustments to optimize the performance of both machine learning methods (Kavzoglu Teke, 2022). While the findings contribute significantly to the field, it is acknowledged that reference support in the discussion is limited. To strengthen the study, additional references, particularly from previous research on machine learning applications in oceanography or similar domains, should be incorporated (Rubbens *et al*., 2023). This would provide a more robust foundation for the discussed methodologies and results, enhancing the credibility

and depth of the research. The study lays the groundwork for continued development in understanding and predicting ocean temperature dynamics in the specific regions under consideration.

Conclusion

This study compares the XGBoost and Random Forest algorithms in predicting the temperature of ocean currents in the Gulf Stream and Labrador Current using data from NASA. The results show that XGBoost has higher accuracy than Random Forest. Visualization of the predictions showed that XGBoost produced highly accurate temperature predictions, indicated by the prediction points being close to the actual temperature reference line. Further statistical analysis confirmed that XGBoost provided more accurate results with a low MSE value (0.228929) and a high R-squared value (0.992583). In the comparison between the two models, XGBoost was superior in prediction accuracy to Random Forest, which had a higher MSE value (0.601143) and lower R-squared value (0.95681). Model evaluation showed that XGBoost was more effective in explaining ocean temperature variations based on historical data, indicating superior performance in ocean temperature prediction. This study suggests the need for further evaluation and parameter adjustment to optimize the performance of both models. The findings make an important contribution to the understanding and prediction of ocean temperature dynamics in the Gulf Stream and Labrador Current region, and pave the way for further development in this area.

References

Asselman, A., Khaldi, M. & Aammou, S. 2023. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interact. Learn. Environ.*, 31(6): 3360–3379. https://doi.org/10.1080/10494820.2021.1928235

Azari, B., Hassan, K., Pierce, J. & Ebrahimi, S. 2022. Evaluation of machine learning methods application in temperature prediction. *Environ. Eng.*, 8(1): 1–12. https://doi.org/10.52547/crpase.8.1.2747

Birant, D.E.R.Y.A. 2011. Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models. *J. Environ. Inform.*, 17(1). https://doi.org/10.3808/jei.201100186

Board, O.P. 2021. Labrador Shelf Offshore Area Strategic Environmental Assessment Update. *Aivek Stantec Limited Partnership*.

Bruera, R., Curbelo, J., García-Sánchez, G. & Mancho, A.M. 2023. Mixing and geometry in the north atlantic meridional overturning circulation. *Geophys. Res. Lett.*, 50(7): e2022GL102244. https://doi.org/10.1029 /2022GL102244

Capotondi, A., Jacox, M., Bowler, C., Kavanaugh, M., Lehodey, P., Barrie, D., Brodie, S., Chaffron, S., Cheng, W., Dias, D.F. & Eveillard, D. 2019. Observational needs supporting marine ecosystems modeling and forecasting: From the global ocean to regional and coastal systems. *Frontiers in Mar. Sci.*, *6*: 623. https://doi.org/10.3389 /fmars.2019.00623

Chen, X. 2022. Effects of Global Climate Changes on Marine Fishery Resources. In Theory and Method of Fisheries Forecasting. Singapore: *Springer Nature Singapore*, p.173–199. https://doi.org/10.1007/978-981-19-2956-4_7

Cui, H., Tang, D., Liu, H., Liu, H., Sui, Y., Lai, Y. & Gu, X. 2024. Modelling Ocean Cooling Induced by Tropical Cyclone Wind Pump Using Explainable Machine Learning Framework. *IEEE Transactions on Geoscience and Remote Sensing*. https://doi.org/10.1109/TGRS.2024.3358374

Devos, L., Meert, W. & Davis, J. 2020. Fast gradient boosting decision trees with bit-level data structures. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany. *Springer International Publishing*, p.590–606. https://doi.org/10.1007/978-3-030-46150-8_35

Duan, S.B., Lian, Y., Zhao, E., Chen, H., Han, W. & Wu, Z. 2023. A Novel Approach to All-Weather LST Estimation using XGBoost Model and Multi-Source Data. *IEEE Transactions on Geoscience and Remote Sensing*. https://doi.org/10.1109/TGRS. 2023.3324481

Ferchichi, H., St-Hilaire, A., Ouarda, T. B. & Lévesque, B. 2022. Prediction of coastal water temperature using statistical models. *Estuar. Coasts*, 45(7): 1909–1927. https://doi.org/10.1007/s12237-022-01070-0

Gonçalves Neto, A., Palter, J.B., Xu, X. & Fratantoni, P. 2023. Temporal Variability of the Labrador Current Pathways Around the Tail of the Grand Banks at Intermediate Depths in a High-Resolution Ocean Circulation Model. *J. Geophys. Res. Oceans*, 128(3): e2022JC018756. https://doi.org/ 10.1029/2022JC018756

He, J., Hao, Y. & Wang, X. 2021. An interpretable aid decision-making model for flag state control ship

detention based on SMOTE and XGBoost. *J. Mar. Sci. Eng.*, 9(2): p.156. https://doi.org/10.33 90/jmse9020156

Hodson, T.O. 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev.*, 15(14): 1-10. https://doi.org/10.5194/ gmd-2022-64

Hussain, M.A., Chen, Z., Wang, R. & Shoaib, M. 2021. PS-InSAR-based validated landslide susceptibility mapping along Karakorum Highway, Pakistan. *Remote Sens.*, 13(20): p.4129. https://doi.org/10.3390/rs132041 29

Ishwaran, H. 2015. The effect of splitting on random forests. *Machine Learning*, 99: 75–118. https://doi.org/10.1007/s10994-014 -5451-2

Jacobs, Z.L., Yool, A., Jebri, F., Srokosz, M., van Gennip, S., Kelly, S.J., Roberts, M., Sauer, W., Queiros, A.M., Osuka, K.E. & Samoilys, M. 2021. Key climate change stressors of marine ecosystems along the path of the East African coastal current. *Ocean Coast. Manag.*, 208: p.105627. https://doi.org/10.1016/j.ocecoam an.2021.105627

Kavzoglu, T. & Teke, A. 2022. Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). *Bull. Eng. Geol. Environ.*, 81(5): p.201. https:// doi.org/10.1007/s10064-022-02708-w

Khan, N., Shahid, S., Ismail, T. B. & Behlil, F. 2021. Prediction of heat waves over Pakistan using support vector machine algorithm in the context of climate change. *Stoch. Env. Res. Risk A.*, 35: 1335–1353. https://doi.org/10.1007/s004 77-020-01963-1

Liu, Z., Wen, T., Sun, W. & Zhang, Q. 2020. Semi-supervised self-training feature weighted clustering decision tree and random forest. *IEEE Access*, 8: 128337–128348. https://doi.org /10.1109/ACCESS.2020.3008951

Lochte, A.A., Repschläger, J., Seidenkrantz, M.S., Kienast, M., Blanz, T. & Schneider, R.R. 2019. Holocene water mass changes in the Labrador Current. *The Holocene*, 29(4): 676–690. https://doi.org/10.1177/0959683618824752

Moraga, P. & Baker, L. 2022. rspatialdata: A collection of data sources and tutorials on downloading and visualising spatial data using R. *F1000Res.*, *11*. https://doi.org/10. 12688/ f1000research.122764.1

Naser, M.Z. & Alavi, A. 2020. Insights into performance fitness and error metrics for machine learning. *arXiv Preprint arXiv:2006.00887*. https://doi.org/10.485 50/ arXiv.2006.00887

Oukawa, G.Y., Krecl, P. & Targino, A.C. 2022. Fine-scale modeling of the urban heat island: A comparison of multiple linear regression and random forest approaches. *Sci. Total Environ.*, 815: 152836. https://doi.org/10.1016/j.scito tenv.2021.152836

Pannozzo, L. 2023. The devil and the deep blue sea: An investigation into the scapegoating of Canada's grey seal. Canada: Fernwood Publishing.

Qiu, Y., Zhou, J., Khandelwal, M., Yang, H., Yang, P. & Li, C. 2022. Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Engineering with Computers*, 38(Suppl 5): 4145–4162. https://doi.org/10. 1007/s00366-021-013 93-9

Reddy, G.T., Bhattacharya, S., Ramakrishnan, S.S., Chowdhary, C.L., Hakak, S., Kaluri, R. & Reddy, M.P.K. 2020. An ensemble-based machine learning model for diabetic retinopathy classification. ic-ETITE, p.1–6. https://doi.org/ 10.1109/icETITE47903.2020.235

Rong, G., Alu, S., Li, K., Su, Y., Zhang, J., Zhang, Y. & Li, T. 2020. Rainfall induced landslide susceptibility mapping based on Bayesian optimized random forest and gradient boosting decision tree models—A case study of Shuicheng County, China. *Water*, 12(11): p3066. https://doi.org/10.3390/w121130 66

Rubbens, P., Brodie, S., Cordier, T., Destro Barcellos, D., Devos, P., Fernandes-Salvador, J.A., Fincham, J.I., Gomes, A., Handegard, N.O., Howell, K. & Jamet, C. 2023. Machine learning in marine ecology: An overview of techniques and applications. *ICES J. Mar. Sci.*, 80(7): 1829–1853. https://doi.org/10.1093/icesj ms/fsad100

Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K. & Ramachandran, M. 2021. A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. *Process.*, 9(11): p.2015. https://doi.org/10 .3390/pr9112015

Shokri, R., Stronati, M., Song, C. & Shmatikov, V. 2017. Membership inference attacks against machine learning models. In 2017 IEEE

symposium on security and privacy (SP), p.3–18. https://doi.org/10 .1109/SP.2017.41

Song, Y.Y. & Ying, L.U. 2015. Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2): 130. https://doi.org/10.11919/j.issn.1002-0829. 215044

Trossman, D. & Palter, J. 2021. Changing ocean currents. From Hurricanes to Epidemics: The Ocean's Evolving Impact on Human Health-Perspectives from the US. Springer, p.11–26. https://doi.org/10.1007/978-3-030-55012-7_2

Urban, M.C. 2019. Projecting biological impacts from climate change like a climate scientist. *Wiley Interdiscip. Rev. Clim. Change*, 10(4): p.585. https://doi.org/10.1002/wcc.585

Wang, Y. & Guo, Y. 2020. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. *China Commun.*, 17(3): 205–221. https://doi.org/10. 23919/JCC.2020.03.017

Wei, J., Huang, W., Li, Z., Sun, L., Zhu, X., Yuan, Q., Liu, L. & Cribb, M. 2020. Cloud detection for Landsat imagery by combining the random forest and superpixels extracted via energy-driven sampling segmentation approaches. *Remote Sens. Environ.*, 248: p.112005. https://doi. org/10.1016/j.rse.2020.112005

Williams, D. 2021. Transatlantic climate and Gulf Stream aesthetics. *Nineteenth-Century Literature*, 76(1): 57–91. https://doi.org/10.15 25/ncl.2021.76.1.57

Wolff, S., O'Donncha, F. & Chen, B. 2020. Statistical and machine learning ensemble modelling to forecast sea surface temperature. *J. Mar. Syst.*, 208: p.103347. https://doi.org/10.1016/ j.jmarsys.2020.103347