Sea Surface Temperature Analysis by Month, Latitude, and Year

Wan Dang
vdang20@gmu.edu

05/03/2025

Abstract - In this study, we investigate the relationship between time, geographic latitude, and sea surface temperature through the application of multiple machine learning algorithms. By leveraging structured climate data—including month, year, and latitude as predictors—we evaluate the effectiveness of linear regression, polynomial regression, and random forest models in predicting sea surface temperature. This project represents an effort to quantify and model environmental patterns that may relate to broader climate trends. Through residual analysis, model comparison, and visualizations, we assess the capability of each algorithm to capture seasonal variation and spatial gradients. While the current dataset is temporally imbalanced, our findings suggest that nonlinear models—particularly random forest—are better suited to handle the complexity of climate behavior. This work lays the foundation for further study, especially as part of a broader effort to explore historical warming in the Antarctic using early 20th-century data from Norwegian whaling expeditions [1].

1. INTRODUCTION

In today's world, the signs of climate change are increasingly visible and scientifically undeniable. Global warming trends are accelerating—particularly since the mid-20th century—and the effects are especially pronounced in polar regions. As snow and ice melt in the Arctic and Antarctic, Earth's reflective surface area decreases, amplifying heat absorption and hastening warming through a process known as ice-albedo feedback. Understanding whether the polar regions began warming earlier than the global average is central to long-term climate modeling.

This study focuses on building predictive models to understand how sea surface temperature correlates with date and latitude. Using a dataset composed of time-stamped temperature readings and geolocations, we test three machine learning models—linear regression, polynomial regression, and random forest regression—to determine how well each can model spatial-temporal temperature variation. This report emphasizes both technical accuracy and model interpretability as we work toward answering a deeper climatological question: Did the Antarctic show signs of early warming before the rest of the world? While this report serves as a modeling benchmark, the broader goal is to extend this methodology to historical datasets, such as those collected by Norwegian whaling vessels operating in Antarctic waters during the early 20th century.

2. METHODOLOGIES

2.1 Data Cleaning and Preprocessing

The raw dataset included missing values, unformatted date strings, and inconsistent temporal coverage across seasons. To prepare the data for modeling sea surface temperature (SST), we applied the following cleaning and feature engineering steps:

- Removed records lacking valid sea temperature or latitude values, which are critical for spatial-temporal analysis.
- Converted the `date` column into Python datetime format to enable structured time feature extraction.
- Extracted `month` and `year` from each observation to study seasonal and interannual trends.
- Applied cyclic transformation to month via `month_sin` and `month_cos` to avoid artificial discontinuities between December and January.
- Standardized `year` to have zero mean and unit variance to ensure temporal features contribute evenly to model training.

2.2 Model Selection Rationale

To investigate the relationship between sea surface temperature, time, and latitude, we selected three regression models representing increasing levels of complexity:

1. **Linear Regression:** Establishes a baseline model, enabling interpretation of individual contributions from time and location on temperature. Useful for examining broad warming trends.

2. **Polynomial Regression (degree 2):** Incorporates nonlinear and interaction effects (e.g., how the influence of latitude may change across seasons), helping to identify curvature in SST changes.

3. **Random Forest Regression:** A flexible, nonparametric model capable of learning complex and localized relationships in the data. This is particularly useful for detecting anomalies or abrupt changes in temperature that may precede broader trends.

2.3 Evaluation Metrics

To assess the predictive performance and interpretive utility of each model, we used the following metrics:

- R^2 Score: Measures the proportion of temperature variability explained by the model.

- RMSE (Root Mean Squared Error): Quantifies average prediction error in degrees Celsius, aiding climatological interpretation.

- Residual vs. Fitted Plots: Diagnose model fit and potential systematic biases or seasonally-dependent errors.

- 3D Temperature Surface Visualization: Visual comparison of model predictions against actual SST data across time and latitude.

- Feature Importance (Random Forest): Highlights which variables most influence SST predictions—particularly useful for assessing the role of time vs. geography in historical warming.

3. VISUALIZATIONS

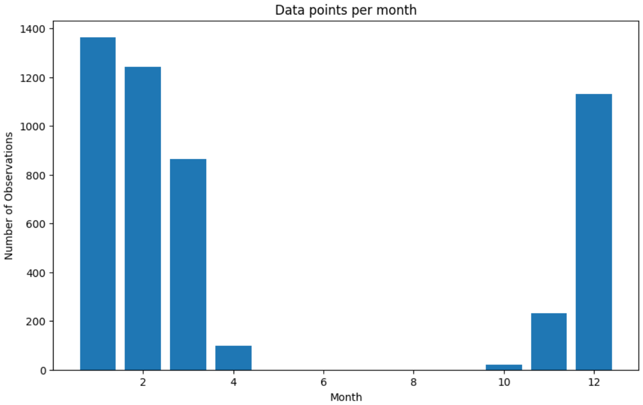

Figure 1: Data Distribution by Month

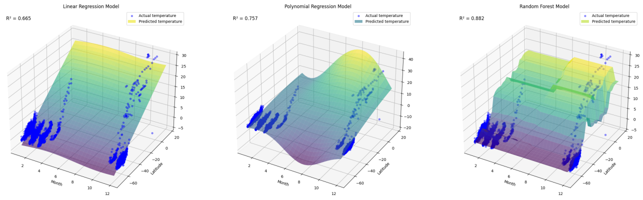

Figure 2: 3D Predicted Surface Comparison Across Models

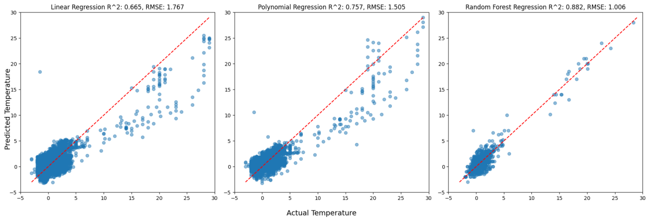

Figure 3: Predicted vs. Actual Temperatures Across Models

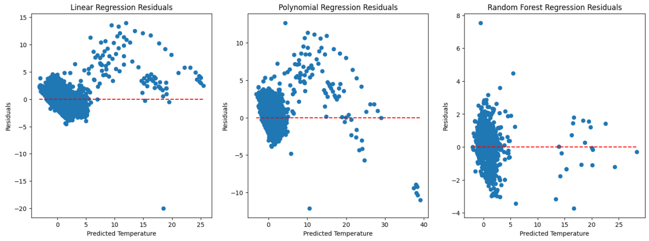

Figure 4: Residuals vs. Predicted Temperature Across Models

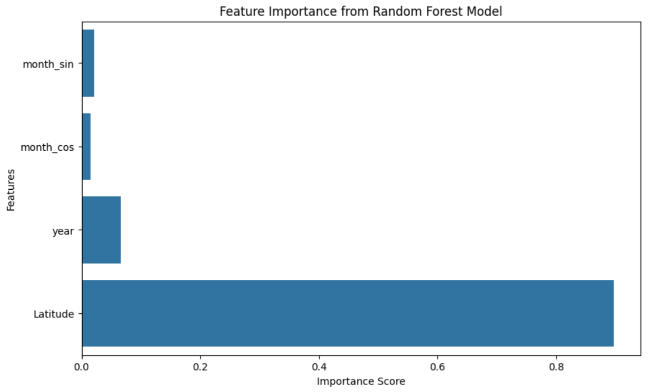

Figure 5: Feature Importance from Random Forest Model

4. DISCUSSIONS

The performance of three regression models - linear regression, polynomial regression, and random forest - was evaluated to predict temperature based on time (month, year) and spatial (latitude) features.

First, the 3D surface plots of predicted temperature Fig 2 illustrate how each model fits the seasonal and spatial structure of the data. The linear regression model produces a smooth planar surface, but failing to fully capture the curvature in observed temperature. In contrast, the polynomial regression model introduces more flexibility and is visibility better at following the underlying shape of the actual temperature values. The random forest model fits the data most tightly, especially around dense regions, producing a stepped surface that captures local variations effectively. These observations align with the respective R^2 scores: 0.665 for linear regression, 0.757 for polynomial regression, and 0.882 for random forest, confirming increasing explanatory power with model complexity.

This improvement is also apparent in the predicted vs actual scatterplot Fig 3. The linear model shows widespread deviation from the red identity line, especially for higher temperature values. The polynomial model aligns more closely but still exhibits some underestimation at high values. The random forest model shows the tightest clustering around the identity line, indicating more accurate predictions across the temperature spectrum. Corresponding RMSE values further reinforce this trend: 1.767 for linear regression, 1.505 for polynomial regression, and just 1.006 for random forest.

Residual analysis provides further insight into each model's limitations Fig 4. The linear regression residuals show a parabolic shape, indicative of nonlinearity and model misfit. The polynomial regression residuals reduce this trend but still show structured error, especially at higher predicted temperatures. The random forest residuals exhibit no apparent trend and are tightly centered around zero, confirming homoscedasticity and well-controlled prediction errors.

However, despite the strong performance of tree-based models, feature importance analysis from the random forest Fig 5 reveals a heavy dependence on latitude, with very low importance scores assigned to month and year. This indicates that the model learned spatial variance well, but was less effective at capturing temporal effects, likely due to limitations in the dataset.

The distribution of data across months confirm this Fig 1. The majority of observations come from winter months (January-March and December), with a dramatic drop-off in summer months (May-September). This temporal imbalance restricts the model's ability to generalize seasonally, particularly when predicting for underrepresented or unseen months. As a result, while the model performs well overall, it may produce unreliable predictions during warmer months or in equatorial regions where seasonal patterns are distinct.

In summary, all models demonstrated improvement with complexity, and the random forest model offered the best overall predictive accuracy. However, the lack of balanced temporal coverage and the dominance of latitude in feature importance suggest the model's learning is biased toward spatial rather than seasonal variation. Future work should use more complete dataset with year-round observations to better isolate and quantify seasonal and temporal effects, including potential warming trends.

5. CONCLUSIONS

This study demonstrates the potential of using machine learning techniques to model sea surface temperature (SST) based on temporal (year and month) and spatial (latitude) features. Among the models tested, random forest regression outperformed both linear and polynomial regression in predictive accuracy, achieving an R^2 score of 0.882 and RMSE of approximately 1.006°C. This suggests that non-linear, ensemble-based models are better suited to capture the complexity of SST dynamics.

However, further analysis revealed that while time and latitude are informative, they alone are insufficient to fully explain SST variability. Feature importance analysis from the random forest model shows that latitude contributed disproportionately to prediction accuracy, while cyclic month and standardized year played relatively minor roles. This likely reflects limitations in the dataset—particularly its seasonal imbalance—as well as the exclusion of other environmental variables that influence ocean temperature, such as ocean currents, salinity, atmospheric conditions, and ice coverage.

To address these limitations and improve future analyses, more comprehensive methods and richer datasets should be considered. Techniques such as Principal Component Analysis (PCA) can help extract dominant modes of variability in high-dimensional environmental datasets, aiding in the separation of natural variability from anthropogenic trends [2]. Additionally, Bayesian hierarchical models offer a powerful framework for quantifying uncertainty and attributing observed changes in temperature to specific drivers over space and time [3]. Furthermore, Geographically Weighted Regression (GWR) enables the modeling of spatially varying relationships, which is particularly relevant when examining regional warming disparities [4]. Finally, nonlinear dimensionality reduction techniques, such as nonlinear PCA, can uncover more complex spatiotemporal patterns that traditional models may miss [5].

While this study provides a first step in correlating SST with time and location, the findings also underscore the importance of incorporating additional environmental variables and advanced modeling techniques. These future directions are essential for uncovering early indicators of global warming—especially in regions such as the Antarctic, where historical whaling data may reveal whether signs of polar warming preceded global trends.

REFERENCES

[1] O. E. Bjørge, K. I. Ugland, and D. Divine, “Whaling statistics, weather and sea ice data from the southern ocean for the period of 1932 to 1963 from catch logbooks of factory ships of company thor dahl a/s.” 2023. [Online]. Available: https://data.npolar.no/dataset/b9f318f5-dda6-444a-9de9-e880f94944f4

[2] Y. Jiang, D. Cooley, and M. F. Wehner, “Principal component analysis for extremes and application to u.s. precipitation,” *Journal of Climate*, vol. 33, no. 15, p. 6441–6451, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1175/JCLI-D-19-0413.1

[3] M. Katzfuss, D. Hammerling, and R. L. Smith, “A bayesian hierarchical model for climate change detection and attribution,” *Geophysical Research Letters*, vol. 44, no. 11, p. 5720–5728, Jun. 2017. [Online]. Available: http://dx.doi.org/10.1002/2017GL073688

[4] H. Han, Z. Zeeshan, B. A. Talpur, T. Sadiq, U. A. Bhatti, E. M. Awwad, M. Al-Razgan, and Y. Y. Ghadi, “Studying long term relationship between carbon emissions, soil, and climate change: Insights from a global earth modeling framework,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 130, p. 103902, Jun. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.jag.2024.103902

[5] D. Bueso, M. Piles, and G. Camps-Valls, “Nonlinear pca for spatio-temporal analysis of earth observation data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, p. 5752–5763, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2020.2969813