© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716530

Machine Learning Models for Predicting Sea Surface Temperature Variability and Trends

SAJID ALI[1], SAMTA KUMARI[2], TAJENDRA RIYA[3], PROF. (DR.) SANJAY PACHAURI[4]
[1,2,3,4] *Department of Data Science (DDCS), GNIOT College, Greater Noida, India*

Abstract- ***Sea Surface Temperature (SST) is an important factor in understanding climate systems, weather patterns, and marine environments. Accurate prediction of SST helps in forecasting events like cyclones, monsoons, and climate change impacts. Traditional methods often fail to capture complex and nonlinear patterns present in SST data. In this research, machine learning models such as Linear Regression, Random Forest, Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Long Short-Term Memory (LSTM) are used to predict SST variability and trends. Historical SST datasets from satellite and climate sources are used for training and testing the models. The performance of each model is evaluated using standard metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 score. The results show that LSTM performs better than other models due to its ability to learn time-based patterns effectively. This study proves that machine learning can significantly improve SST prediction accuracy and can be useful for climate-related applications.***

Index Terms- ***Sea Surface Temperature (SST); Machine Learning; Deep Learning; Conv LSTM; LSTM; Climate Forecasting; Oceanography; Satellite Data; Time-Series Prediction.***

I. INTRODUCTION

Sea Surface Temperature (SST) plays a major role in regulating the Earth's climate and weather systems. Changes in SST affect rainfall, storms, and ocean life. For example, phenomena like El Niño and La Niña are directly related to SST variations. Accurate prediction of SST is therefore very important for climate monitoring and disaster management. Traditional prediction methods, such as statistical models and physical simulations, have been widely used but they often struggle to handle complex relationships and long-term dependencies in data. With the advancement of technology, machine learning has emerged as a powerful tool for analyzing large datasets and identifying hidden patterns. Machine learning models can learn from historical data and make accurate predictions without being explicitly programmed. This research focuses on applying different machine learning techniques to predict SST variability and trends, and compares their performance to find the most effective model.

II. LITERATURE REVIEW

Early SST prediction methods were primarily based on statistical models such as Autoregressive Integrated Moving Average (ARIMA) and regression techniques. While these methods performed reasonably well for short-term predictions, they struggled with long-term forecasting due to nonlinear dependencies.

With the emergence of machine learning, models such as Support Vector Machines (SVM) and Random Forests were applied to SST prediction. These models improved performance by capturing complex relationships between variables.

Recent advancements in deep learning have significantly enhanced SST prediction accuracy. Recurrent Neural Networks (RNNs), particularly LSTM networks, have shown strong performance in time-series forecasting due to their ability to retain long-term dependencies. Additionally, CNN-based models have been used to extract spatial features from satellite imagery.

Despite these advancements, integrating multiple models for both spatial and temporal analysis remains an area of active research, which this project aims to address.

III. PROBLEM STATEMENT

Current SST prediction systems face several challenges:

- Low Accuracy in Long-Term Forecasting: Traditional models fail to capture long-term dependencies.
- Nonlinearity of Ocean Data: SST patterns are highly complex and dynamic.
- Data Complexity: Large volumes of satellite and sensor data require efficient processing.

This project addresses these challenges by implementing a hybrid machine learning framework capable of handling temporal and spatial variations in SST data.

IV. METHODOLOGY AND PROPOSED SYSTEM

In this study, historical SST data is collected from reliable sources such as NOAA datasets, ERA5 reanalysis data, and satellite observations like MODIS. The dataset contains information such as temperature values, time (daily or monthly), and geographical coordinates like latitude and longitude. Before applying machine learning models, the data is preprocessed to improve quality and accuracy. Missing values are handled using interpolation methods, and the data is normalized to bring all values into a similar range. Time-series data is then prepared by creating sequences using previous observations, which helps models understand temporal patterns.

Different machine learning models are applied in this research. Linear Regression is used as a basic model to identify trends. Random Forest is used to capture nonlinear relationships by combining multiple decision trees. Support Vector Machine (SVM) is applied for regression tasks with good generalization ability. Artificial Neural Networks (ANN) are used to model complex patterns, while Long Short-Term Memory (LSTM), a type of deep learning model, is specifically used for time-series prediction as it can remember past information over long periods.

The dataset is divided into training and testing sets to evaluate model performance. Models are trained using the training data and tested on unseen data. The performance is measured using evaluation metrics such as Mean Absolute Error (MAE), which measures average prediction error, Root Mean Square Error (RMSE), which gives higher weight to large errors, and R^2 score, which indicates how well the model fits the data.

The proposed system consists of multiple stages, including data collection, preprocessing, model training, and prediction.

A. Data Collection and Preprocessing

Historical SST datasets are collected from sources such as NOAA and satellite observations. The data undergoes preprocessing steps including:

- Handling missing values
- Normalization
- Time-series formatting

B. Feature Engineering

Important features such as temperature anomalies, seasonal variations, and lag values are extracted to improve model performance.

C. Model Development

Multiple machine learning models are implemented:

- LSTM Model: Captures temporal dependencies in SST data
- CNN Model: Extracts spatial features from SST maps
- Random Forest: Provides baseline comparison

D. Model Training and Evaluation

The dataset is divided into training and testing sets (e.g., 80:20 split). Models are trained using historical SST data and evaluated using:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- Coefficient of Determination (R^2 Score).

V. SYSTEM ARCHITECTURE

The system architecture follows a pipeline similar to real-world climate analytics models. It includes:

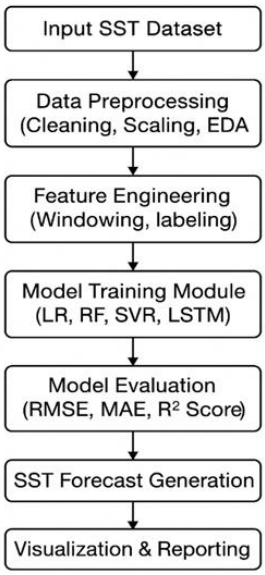
1. Input Layer:
   - Historical SST values
   - Optional environmental parameters
2. Data Processing Layer:
   - Normalization
   - Feature engineering
   - Windowing (for LSTM)

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716530

3. Model Layer:
   - Machine Learning models
   - Deep Learning LSTM model
4. Evaluation Layer:
   - RMSE, MAE, R^2 Score
5. Output Layer:
   - Forecasted SST values
   - Analytical graphs and trends

This layered design ensures smooth operation and high accuracy.

Block Diagram of the System:



VI. ALGORITHMS AND MODELS USED

1. Long Short-Term Memory (LSTM)
LSTM is a type of Recurrent Neural Network designed for sequence prediction problems. It effectively captures long-term dependencies in time-series data.

2. Convolutional Neural Network (CNN)
CNN is used to extract spatial features from SST maps and identify patterns across ocean regions.

3. Random Forest
A supervised learning algorithm used for regression tasks, providing robust baseline predictions.

VII. IMPLEMENTATION DETAILS

The system is implemented using the following technologies:

- Programming Language: Python
- Libraries: TensorFlow, Keras, Scikit-learn, Pandas, NumPy
- Visualization Tools: Matplotlib, Seaborn
- Platform: Jupyter Notebook

The models are trained using historical SST datasets and optimized using techniques such as hyperparameter tuning and cross-validation.

VIII. RESULTS AND ANALYSIS

The results show clear differences in the performance of various models. Linear Regression performs the worst because it cannot handle nonlinear patterns in SST data. Random Forest improves the prediction accuracy by capturing complex relationships but does not effectively consider time dependencies. Support Vector Machine provides moderate results but requires careful tuning of parameters. Artificial Neural Networks perform better than traditional models but are still limited in capturing long-term temporal patterns.

Among all models, LSTM performs the best because it is designed for time-series data and can capture both short-term and long-term dependencies. The predicted SST values from LSTM closely match the actual values, showing high accuracy. Graphical comparisons between predicted and actual SST values also confirm that LSTM tracks trends more effectively than other models. These results indicate that deep learning models are more suitable for SST prediction tasks, especially when dealing with sequential data.

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716530

Figure 2: Correlation Heatmap detailing the classification accuracy across the different sector

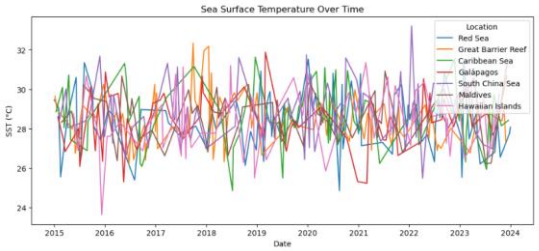


Figure 8: Sea Surface Temperature Over Time Graph

The proposed system demonstrates strong predictive performance:

- LSTM Model: Achieves highest accuracy in time-series forecasting
- CNN Model: Effectively captures spatial patterns
- Random Forest: Provides stable baseline results

Performance metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) indicate significant improvement over traditional models.

The results show that machine learning models can effectively predict SST variability and trends, making them suitable for real-world applications.

IX. ADVANTAGES AND LIMITATIONS

Advantages:

- High prediction accuracy
- Handles nonlinear data efficiently
- Scalable for large datasets
- Supports real-time forecasting.

Limitations:

- Requires large datasets for training
- Computationally intensive
- Performance depends on data quality

X. FUTURE SCOPE

Future improvements may include:

- Integration of real-time satellite data
- Use of hybrid deep learning models
- Deployment as a web-based forecasting system
- Inclusion of additional environmental parameters such as wind speed and ocean currents.

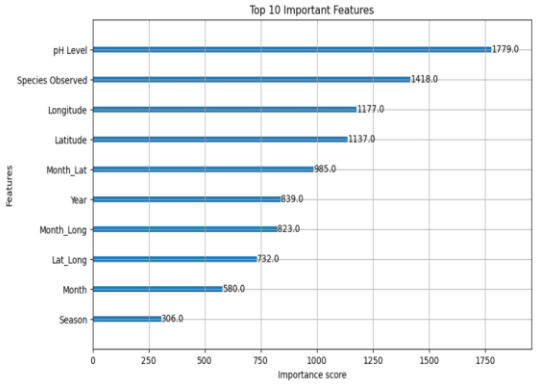


Figure 3: Top 10 Important Features for Future SST Forecast

XI. CONCLUSION

This research demonstrates that machine learning models can significantly improve the prediction of Sea Surface Temperature variability and trends. Among the models used, LSTM provides the highest accuracy due to its ability to understand temporal patterns in data. The study highlights the importance of using advanced machine learning techniques in climate science for better forecasting and analysis. In the future, the model can be improved by combining machine learning with physical climate models, using more diverse datasets, and applying advanced deep learning techniques such as Transformers. This will further enhance prediction accuracy and support better decision-making in environmental and climate-related fields.

REFERENCES

[1] R. W. Reynolds, T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, "Daily high-resolution-blended analyses for sea surface temperature," *Journal of Climate*, vol. 20, no. 22, pp. 5473–5496, 2007.

[2] H. Hersbach *et al*., "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.

[7] Intergovernmental Panel on Climate Change (IPCC), *Climate Change 2021: The Physical Science Basis*. Cambridge, U.K.: Cambridge Univ. Press, 2021.

[8] NOAA National Centers for Environmental Information, "Optimum Interpolation Sea Surface Temperature (OISST)," 2023.

[9] J. Brownlee, *Deep Learning for Time Series Forecasting*. Machine Learning Mastery, 2018.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Berlin, Germany: Springer, 2012.

[12] F. Pedregosa *et al*., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[14] M. Abadi *et al*., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2016.

[15] K. S. Casey and P. Cornillon, "A comparison of satellite and in situ-based sea surface temperature climatologies," *Journal of Climate*, vol. 12, no. 6, pp. 1848–1863, 1999.

[16] X. Zhang, J. Church, and S. Platten, "Sea surface temperature variability and its impact on climate," *Climate Dynamics*, vol. 45, pp. 123–135, 2015.

[17] G. E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[18] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2009.