Project Title: Forecasting of Sea Surface Temperature and Marine Weather Variables using NOAA ICOADS Time Series Data

Umaeshwer Shankar, Arin Paul, Alex Zhang, Ketan Totlani

Project Idea:

The primary aim of this project is to analyze and forecast Sea Surface Temperature (SST) along with other essential marine weather variables, such as air temperature, wind speed, and atmospheric pressure, using the NOAA International Comprehensive Ocean-Atmosphere Data Set (ICOADS)**.** By utilizing time series forecasting techniques, including statistical algorithms like SARIMA, machine learning models such as LSTM, or using a pretrained model like Prophet, the project seeks to build predictive models capable of estimating future oceanic and atmospheric conditions with high accuracy.

This project is motivated by the growing importance of data-driven climate analytics and the availability of rich historical datasets such as NOAA ICOADS, which provide an excellent foundation for developing reliable predictive models. Oceans play a critical role in regulating global climate and accurate forecasting of sea surface temperature and related weather parameters is vital for climate monitoring, marine ecosystem management, navigation safety, and disaster preparedness.

By the end of this project, we aim to develop a robust forecasting framework capable of predicting future sea surface temperatures and related marine weather variables. The project will also focus on identifying seasonal trends, detecting anomalies, and analyzing correlations among various oceanic and atmospheric parameters. Furthermore, it will demonstrate the effectiveness of time series models such as SARIMA, LSTM, and Prophet for forecasting. The outcomes will include comprehensive visualizations and evaluation metrics that clearly illustrate the models' performance and predictive reliability. [1]

Description of Dataset:

The dataset used in this project is the NOAA International Comprehensive Ocean Atmosphere Data Set (ICOADS), publicly available through Kaggle. Data is compiled, maintained, and standardized by the National Oceanic and Atmospheric Administration (NOAA), ensuring reliability and consistency across sources. The ICOADS dataset contains over 800 million records spanning from the mid-1600s to 2017. For computational feasibility, this project will use a filtered subset of the dataset starting from 2001 to 2017.

This project focuses on forecasting Sea Surface Temperature (SST) and related marine weather variables using historical marine weather data. For this, we require long-term, continuous, and time-stamped measurements of variables. The selected dataset provides variables such as SST, Air Temperature, Wind Speed and Direction, Atmospheric Pressure, Relative Humidity, Latitude, Longitude, and Time of Observation. The variables mentioned in this dataset will help us build multivariate time series models using SARIMA, Prophet, or LSTM.

Project Plan:

The project will utilize open-source tools commonly used in data science. Python will be the primary programming language, with Google Colab used for implementation and visualization. Pandas and NumPy will handle data processing. Matplotlib and Seaborn will support visual analysis. Scikit-learn and Statsmodels will be applied for machine learning and time series modeling. PyTorch will power deep learning models like LSTM. Prophet will assist in seasonal forecasting, and Git and GitHub will manage version control.

The project will follow a structured, step-by-step methodology to ensure accurate forecasting and meaningful insights. The first step will be Data Collection and Preprocessing, where we access the NOAA ICOADS dataset using Big Query Python client library, filtered for the period 2001 to 2017. The data will be cleaned by handling missing values, outliers, and inconsistencies, then converted into a time series format with appropriate temporal indexing. In the next step, we will perform statistical summaries to understand variable distributions, seasonal trends, correlations, and anomalies in SST and other marine variables. The Feature Engineering will involve creating time-based features such as month, season, and year, generating lag features and moving averages to capture temporal dependencies, and scaling the data for improved model performance.

In the model development stage, the baseline time series models, such as SARIMA will be implemented for univariate SST forecasting, while advanced models like LSTM and Prophet will be used for multivariate and nonlinear forecasting. This will be done by splitting the dataset into training and testing subsets for evaluation. For model evaluation, we will use performance metrics including Root Mean Square Error (RMSE), and the R^2 Score will be used to compare model accuracy, supported by visualizations of predicted versus actual SST values. The visualization step will focus on generating plots that depict temperature trends, anomaly detection, and forecast projections, along with interpreting the findings to highlight key patterns and relationships among marine parameters. [2]

Division of Work:

Team Member	Responsibilities
Umaeshwer Shankar (umaeshwe@usc.edu)	Data Collection, Modelling Approach (SARIMA)
Arin Paul (arinpaul@usc.edu)	Preprocessing, Modelling Approach (LSTM, Prophet)
Alex Zhang (alexyzha@usc.edu)	Exploratory Data Analysis & Feature Engineering
Ketan Totlani (totlani@usc.edu)	Visualization & Reporting

References:

[1] Q. Huang and Z. Cui, "Study on prediction of ocean effective wave height based on hybrid artificial intelligence model," *Ocean Engineering*, vol. 289, Part 1, p. 116137, 2023, doi: 10.1016/j.oceaneng.2023.116137.

[2] E. Popovska and G. Georgieva-Tsaneva, "Comparative Analysis of ARIMA and LSTM Models for Seasonal Times-Series Forecasting," *2024 15th National Conference with International Participation (ELECTRONICA)*, Sofia, Bulgaria, 2024, pp. 1-3, doi: 10.1109/ELECTRONICA63645.2024.11146090.