Forecasting Sea Surface Temperature and Marine Weather Variables with ICOADS Time Series Data

Umaeshwer Shankar, Arin Paul, Alex Zhang, Ketan Totlani

Project Overview:

The central aim of our project is to leverage the comprehensive NOAA International Comprehensive Ocean-Atmosphere Data Set (ICOADS) to analyze and forecast SST and other essential marine weather variables like air temperature and wind speed. By applying time series forecasting and machine learning techniques to ICOADS, our research seeks to build robust predictive models that can estimate future oceanic and atmospheric conditions with high precision. This includes detecting seasonal trends, identifying anomalies, and visualizing complex relationships among marine variables, which are all crucial steps for advancing risk management and sustainable decision-making in the maritime sector.

Project Plan:

For advanced modeling, our analysis focuses on high-quality ICOADS records from 2005 to 2017. The dataset, updated daily and available through Google BigQuery, offers seamless integration with data science tools, supporting efficient conversion to formats compatible with Spark and Pandas. NOAA ensures the data remains reliable and accessible for research purposes.

The workflow begins with data extraction and cleaning of NOAA ICOADS records (2005–2017) via the BigQuery Python client. After preprocessing, statistical analysis examines distributions, trends, and anomalies in sea surface temperature and related variables. Feature engineering generates relevant time-based and lag features, and prepares the data for modeling. Models such as SARIMA, LSTM, and Prophet are then trained and validated using metrics like RMSE and R^2 Score, with visualizations aiding the comparison of predictions versus actual results. The final stage interprets these outputs, identifying key patterns and relationships among marine parameters.

Division of Work:

Team Member	Responsibilities
Umaeshwer Shankar (umaeshwe@usc.edu)	Data Aggregation, Statistical Analysis
Arin Paul (arinpaul@usc.edu)	Modelling Approaches (LSTM, Prophet)
Alex Zhang (alexyzha@usc.edu)	Cleaning, Exploratory Data Analysis & Feature Engineering
Ketan Totlani (totlani@usc.edu)	Documentation

Updates and Modifications in Plan:

Our exploratory data analysis and data aggregation was done entirely on Google's BigQuery platform instead of Google Colab, as originally planned. This change was made in response to compute limitations associated with

the free tier of Google Colab, which can restrict session length and available memory. This was ultimately not compatible with our dataset, which is around 100 GB in size.This switch improves performance when working with large-scale data and positions our project to take full advantage of scalable, cloud-native resources.

Project Progress:

We performed data aggregation from the public data workspace to our project workspace on BigQuery, keeping the year range between 2005 and 2017. We then stored our aggregated data as a partitioned table for improved query performance. After that, we proceeded to the cleaning and exploratory data analysis stage. We first examined the data schema and dropped the columns which are not relevant for our analysis. Then, we proceeded to remove the features which are highly correlated with SST (Figure 1), which included all of the other temperature variables.

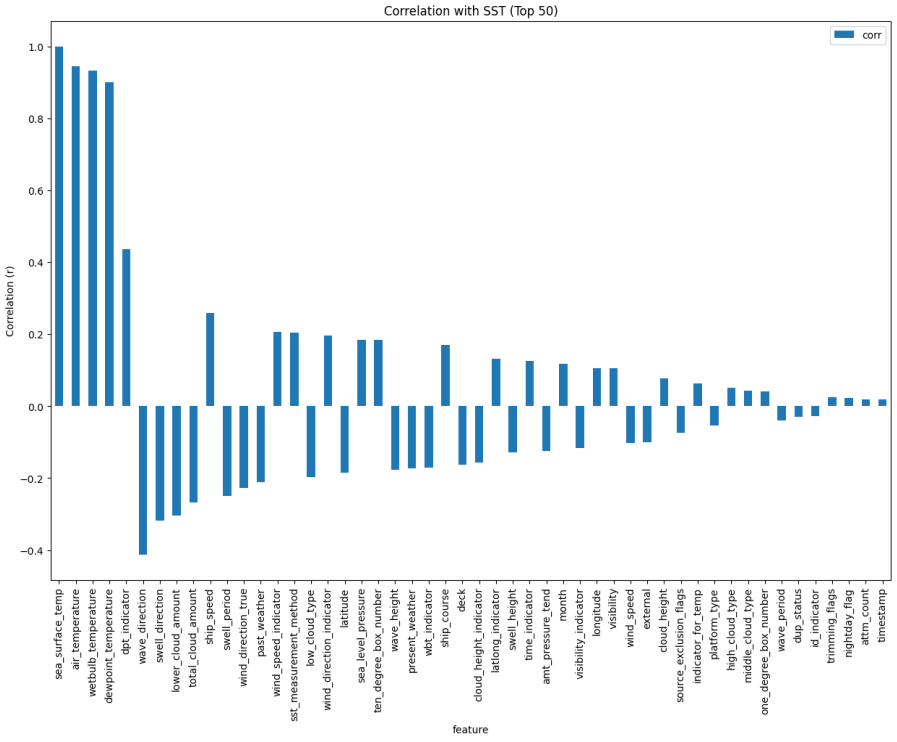

Figure 1: Top 50 variables most correlated (by magnitude) with SST

We then proceeded to identify patterns in the missing data using the pairwise missingness dependence plot (Figure 2) and the missing pattern heatmap (Figure 3). From figure 2, we can see that the missingness of all variables depend on at least 1 other variable. We therefore determined that our data is not missing completely at random. In order to determine whether our data was missing at random, we then generated a missing pattern heatmap, as shown in figure 3. This heatmap was generated with 10 variables of interest, namely: sea surface temperature, sea level pressure, wind speed, as well as some wave-related variables and other temperature variables. From figure 3, we can see that SST being the only recorded variable is the predominant missingness pattern among these variables: there are around 75 million instances of SST being the only recorded variable among these 10. We can also see that, across the most common missing patterns, there is almost always at least one temperature variable present. Additionally, the most commonly recorded variables were SST and SLP (sea level pressure),

as all of the most common missing patterns contained records for SST, SLP, or both variables. Therefore, we conclude that our data isn't MCAR or even fully MAR.

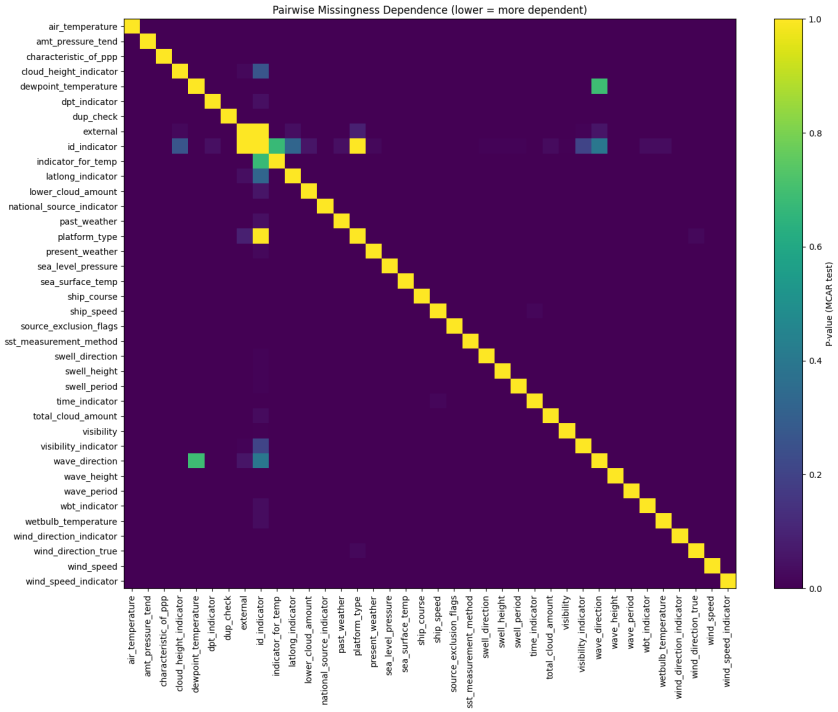

Figure 2: Pairwise MCAR test matrix

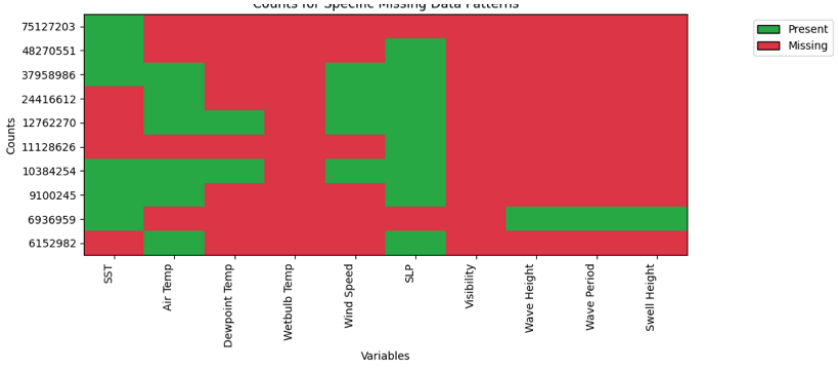

Figure 3: Patterns of Missing Data

After that, we proceeded to create and analyze the time series plots for sea surface temperature (Figures 4-7). Although there doesn't seem to be an overall increase/decrease in SST over time, we can see from the trend and remainder time series plots (Figure 6, 7) that there has been more variance in SST in recent years. We can also see from the seasonal time series plot (Figure 5) that the amplitude of yearly SST recurrences has been increasing in more recent years. This is reflected in the monthly SST mean time series plot (Figure 4) as well, as we can see that years after 2014 see a larger range of mean SSTs compared to years before 2014.

We also performed univariate and bivariate analysis of some variables of interest (SST, air temperature, wind speed, SLP) to determine if they were normally distributed. Through this, we found that our wind speed variable followed a skewed normal distribution (Figure 9).

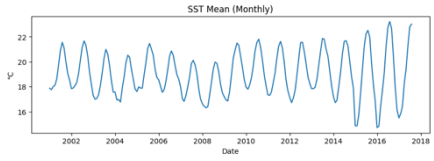

Figure 4: SST Monthly Mean

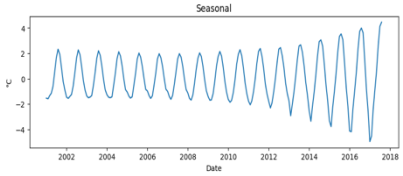

Figure 5: SST Seasonal Plot

Figure 6: SST Remainder Plot

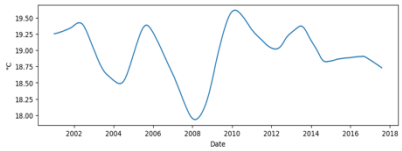

Figure 7: SST Trend Plot

Because of this, we performed a log-normalization transformation to make wind speed normally distributed. We then verified that the transformed distribution was consistent with our previous correlation plot (Figure 10). As seen in

Figure 10 (left), SST and air temperature are strongly correlated with each other, while wind speed and SLP are not correlated with any of the present variables. We can see that this still holds true in figure 10 (right). Additionally, wind speed pre-transformation is now strongly correlated with wind speed post-transformation, meaning that our transformation was valid.

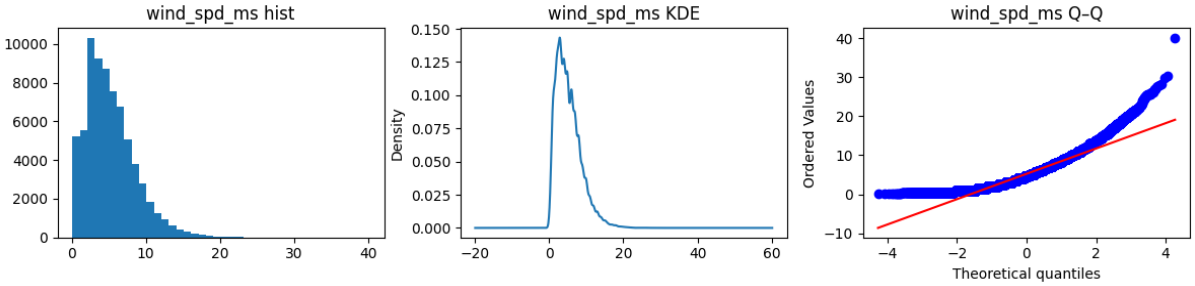

Figure 9: Distribution and QQ Plot for Wind Speed

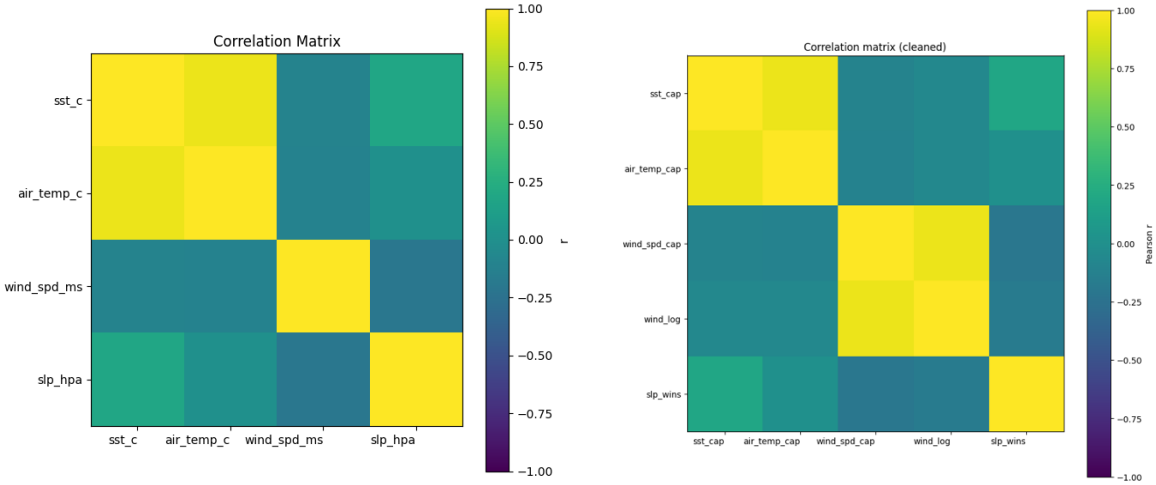

Figure 10: Correlation matrix between variables before (left) and after (right) log-transformation

After the completion of data preparation and analysis, we are now in the stage of forecasting methods. We are currently on statistical methods like Auto-Regressive Integrated Moving Average (ARIMA), which use lag functions and repeated differencing to forecast SST values. The major benefit of this method is that it is very useful when other environmental variables are sparsely populated, which has been observed in our case.

References:

[1] "Global Maritime Freight Transport Market | 2022 - 27 | Industry Share, Size, Growth - Mordor Intelligence," www.mordorintelligence.com. https://www.mordorintelligence.com/industry-reports/global-maritime-freight-transport-market

[2] K. Rjumina, "Climate change, severe weather, impact on shipping | Britannia P&I," Britannia, Jan. 29, 2025. https://britanniapandi.com/2025/01/climate-change-climate-change-severe-weather-and-its-impact-on-shipping-risks/