Prediction of the sea surface temperature in the Pacific Ocean with timeseries Meta model and ARIMA

Contents

Introduction

According to the National Ocean Service (NOAA) (2016), the ocean covers 71 percent of the Earth's surface, and therefore the sea surface temperature (SST), is essential in predicting the weather and for atmospheric model simulations that help study the marine ecosystems. Predicting the SST is also needed in understanding the cycles of the El Nino and La Nina, where "during El Niño, temperatures in the Pacific near the equator are warmer than normal. During La Niña, the same area experiences colder than normal ocean temperatures.

According to the Met Office, machine learning is already being used in the weather production schedule, however the increase in data and need for accuracy pushes the world to constantly search for the best solutions. Met Office highlights that machine learning models are fast and cost significantly less than physics-based simulations, which improved development opportunities, such as larger and higher resolution forecasts with better prediction for extreme weather events, more user focused services and ability to forecast specific scenarios according to user needs.

Importance & Relevance

According to Taccari (2023), the numerical weather prediction methods (NWP) are computationally demanding and tend to come with an accuracy loss. The author continues to highlight that "the last few years have witnessed a shift toward the adoption of data-driven approaches, particularly artificial intelligence, in the realm of weather forecasting. AI integration promises faster models, with training being resource-intensive but only done once. Once trained, these models can analyse new data in real or near-real-time, enabling almost instantaneous predictions. They also have the potential to become sharper and more refined, improving forecast accuracy by identifying complex patterns not easily captured by explicit equations". Dr Mariana Clare from the European Centre for Medium-Range Weather Forecasts, believes that AI is able to accelerate model development whilst maintaining a high level of detail, unlike previous forecasting methods. Taccari also highlights that AI can reveal complex

patterns from the hundreds of millions of daily observations in her research of ground water for weather prediction.

Based on the above, it seems that artificial intelligence and machine learning algorithms are increasing in popularity when it comes to weather prediction due to them being computationally cheap and faster with a higher degree of accuracy and attention to detail. But why such attention to detail is needed when it comes to weather? According to the Met Office (2023), weather prediction is used not only for daily use, but also for policy generation, emergency responders, and business to prepare for high risk natural hazards, such as floods, tornadoes, extreme heatwaves etc. Based on Newman and Noy's article (2023), "the World Meteorological Organization reports that there has been a sevenfold increase in the reported disaster losses from extreme weather since the 1970s" where the economic loss from a natural disaster such as a flood can stem from destroyed housing and roads, or lost crops but also from "declines in economic value-added because of the direct economic damage. The study estimates that "the global cost of extreme weather events in the past 20 years is at £13m an hour, whilst UNICEF recently reported that extreme weather displaced 43m children over the past six years".

Data Source

The data used in this report comes from the UC Irvine Machine Learning Repository and can be found here: https://archive.ics.uci.edu/dataset/122/el+nino. According to the source, the data is collected with the Tropical Atmosphere Ocean (TAO) array which consists of 70 moored buoys spanning the equatorial Pacific. They measure "oceanographic and surface meteorological variables for improved detection understanding and prediction of seasonal-to-interannual climate variations originating in the tropics, most notably those related to the El Nino/Southern Oscillation (ENSO) cycles. [...] Each mooring measures air temperature, relative humidity, surface winds, sea surface temperatures and subsurface temperatures down to a depth of 500 meters and a few a of the buoys measure currents, rainfall and solar radiation."

The data columns in this dataset are: obs, year, month, day, date, latitude, longitude, zon_winds (zonal winds west<0, east>0), mer_winds(meridional winds, south<0, north>0), humidity, air_temp and ss_temp (sea surface temperature – *the target variable* predicted in this report).

The timeline of the data collection varies significantly. The source shows that "data taken from the buoys from as early as 1980 for some locations", which shows there is quite a significant amount of missing data that this report dealt with. Additionally, "the latitude and longitude in the data showed that the buoys moved around to different locations. The latitude values stayed within a degree from the approximate location. Yet the longitude values were sometimes as far as five degrees off of the approximate location.

Ethical considerations

Most of the time, ethical considerations represent the privacy and rights of individuals, however they can apply in environmental modelling as well. One should consider the accuracy and the reliability of the sourced data. In the case of this report, the buoys that collected the data had several gaps with no data (for example on weekends), and their location has fluctuated in various degrees based on the waves and winds. Equally, some of the variables register collected data from 1980, and yet some of them have their data collected at a later date. These aspects need to be brought to consideration as these gaps can introduced bias and poor quality data in models, and have some models perform unexpectedly.

Limitations

This report came with limitations. Firstly, the gap in the data was at times difficult to handle, with data being collected in sporadic years and days. Secondly, due to the nature of the buoys in water, the data is at times collected from slightly different spots, which could perhaps confuse future models. Thirdly, the entire dataset contains roughly 170.000 entries, with latitudes and longitudes from the entire globe, therefore it posed a challenge for this report to narrow down an area for predictions. The combination between the three limitations has made the trends and seasonality of the data be skewed at times.

Methodology

In terms of data distribution, the histogram in the Jupyter Notebook shows that most of the temperature lays between 29.0 and 30.0 degrees Celsius, with the bigger dip in temperature below 28.0 degrees in 1992. Most of the months retain the temperature average between 28.5 and 29.5 degrees with a few outliers in April and in September.

Data cleaning & grouping

This report was initially created on the entire dataset, however due to complications with data collection it made sense to create a smaller dataset based only the longitude with the highest number of value counts, in this case longitude = 165.00.Therefore, the newly created dataframe called elnino_165 was used for data splits and modelling. The 165 longitude represents a spot in the Pacific Ocean.

From a machine learning point, it makes sense to choose one specific geographical point rather than feed the algorithm all the data points from across the globe. This way it can learn a pattern of when it is cold and when it is warm just for one location, instead of getting confused by doing it globally.

Trend of missing values

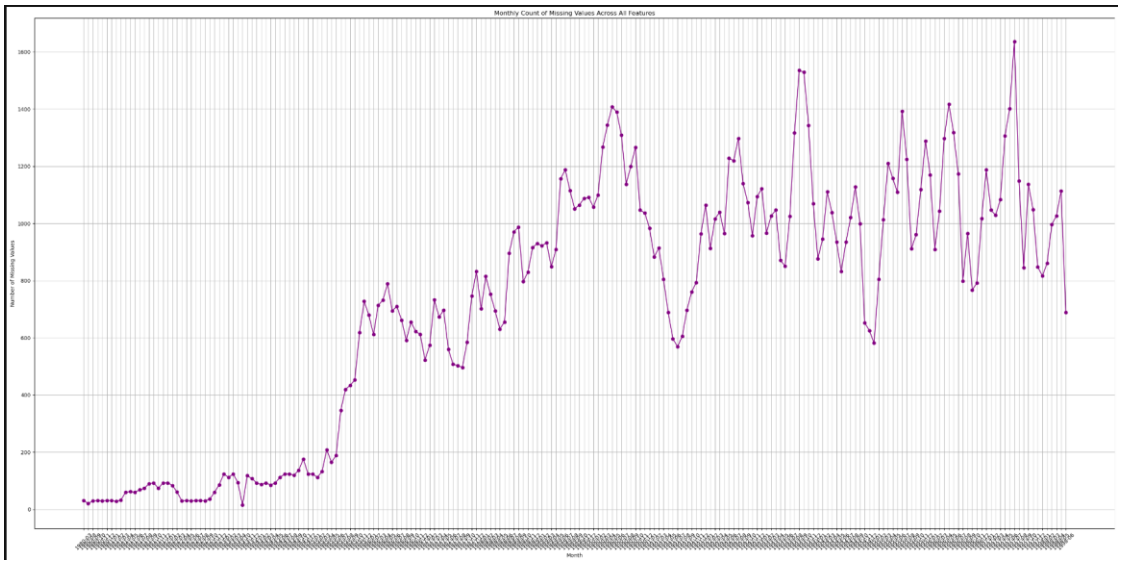


Figure 1. Trend of missing values across all years and all variables

The trend of missing data highlights a steady trend at the start of 1980 and the peak in 1998. The highest number of missing data is over 3000. This is most likely due to the collection of data through buoys launched in water, which can cause gaps due to wind or waves.

Linear interpolation

In order to resolve the trend of missing data, this report looked at linear interpolation. According to "Practical Time Series Analysis" (p 48), "interpolation is a method of determining the values of missing data points based on geometric constraints regarding how we want the overall data to behave. A linear interpolation constrains the missing data to a linear fit consistent with known neighboring points". Linear interpolation allows the user to use their knowledge to fill in the missing data. In this report, the linear interpolation is done looking at past data. This report also used the backfill method for a few more missing data that could not be filled in by the linear interpolation due to the fact that the entries did not have any data before or after that could be used for interpolation.

Feature selection

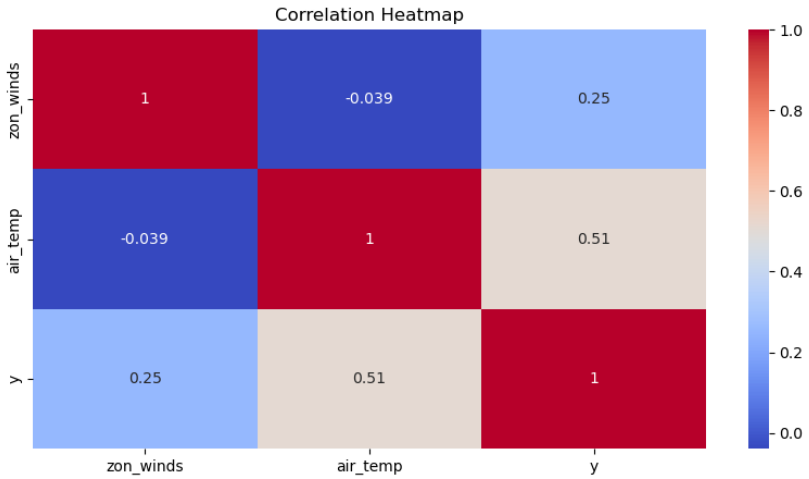


Figure 2 Correlation heatmap for feature selection

The correlation heatmap above shows that the most two features with the highest impact on the prediction of the y variable are the zonal winds and air temperatures. The below image also shows the same two variables in numerical form:

```
#finding out the best regressors for ss_temp
correlations = daily_df.corr()['y'].sort_values(ascending=False)
print(correlations)
```

```
y            1.000000
air_temp     0.509975
zon_winds    0.253377
month        0.221984
Name: y, dtype: float64
```

Figure 3 Feature selection correlation

Working with time series

Stationarity

According to PTSA (p 82), "a stationary time series is one that has fairly stable statistical properties over time, particularity with respect to mean and variance". Also, "a stationary time series is one in which a time series measurement reflects a system in a steady state". An example of data that is not stationary would be data where the mean value is increasing over time, rather than being steady. Equally, when plotted the distance between peaks seems to be growing, therefore the variance is increasing.

This report used a specific function to check if the sea surface temperature data was indeed stationary, and this was the Augmented Dickey-Fuller test. The ADF test is the most common used metric to assess a time series. According to PTSA (p 83), "it posits a null hypothesis that a unit root is present in a time series". The result of the ADF test on the data in this report is that the data is most likely stationary:

```
ADF Statistic: -4.615825
p-value: 0.000121
Critical Values:
        1%: -3.433
        5%: -2.863
        10%: -2.567
The series is likely stationary.
```

Figure 4. ADF test result on sea surface data

Stationarity is necessary in time series analysis firstly because a high number of models assume stationarity, and " a model of a time series that is not stationary will vary in its accuracy as the metrics of the time series vary." Luckily, this was not the case for now.

Seasonality

According to PTSA (p 59), "seasonality is data in any kind of recurring behavior in which the frequency of the behavior is stable. [...] For example, human behavior tends to have a daily seasonality (lunch at the same time every day), a weekly seasonality (Mondays are similar to other Mondays) and a yearly seasonality (New Year's Day has low traffic)".

With that comes the concept of seasonal decomposition which looks at the original time series and decomposes it into a seasonal component, a trend and residuals.

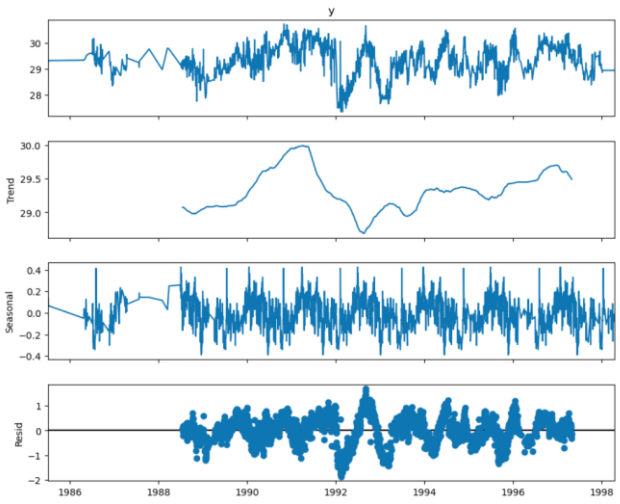


Figure 5. Seasonal decomposition of the sea surface temperature data

The first part represents the observed data, also defined as the "slow-moving changes in a time series" (Time Series Forecasting with Python, p 35). It seems the data has constant ups and downs which confirms the aforementioned aspect of data collection: as the wind and waves took the buoys a way from their initial place combined with the fact that the buoys did not register data on the weekends. The trend component is usually defined as the "level" that "draws a line through most of the data points to show the general direction of a time series" (TSFP , p35). The seasonality component shows that the data is indeed seasonal as it captures seasonal variation, which is " a cycle that occurs over a fixed period of time". (TSFP , p36)

However, it is important to highlight that the seasonality starts roughly around the same time as the trend, which shows the incomplete data collected at first. Finally, the residuals are defined as the "trend and seasonal graphs compared together at each point in time". They show a high variation between the trend and the seasonal graphs, with another start roughly around the same as the other three plots. There is still visual evidence of some identifiable pattern in the residuals plot which could imply that the decomposition is not able to fully identify the seasonality. In order to overcome this challenge, the recommendation would be to finetune the periods in the seasonal decomposition. This period should be the number of data points in a specific cycle, however due to the nature of the data collection identifying these cycles would be challenging.

Models

1. Prophet

Prophet is a is an open source model produced by Meta and according to its own documentation is described as "a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well". It is introduced through the sklearn model API and it contains fit and predict methods. (Prophet, 2023)

The input for Prophet is a dataframe with two columns, ds (date stamp) and y (target). The model is fit on training data and then it creates a dataframe called future which contains the dates for which the prediction is made (which also includes the historical data). Based on the future dataframe, it creates the predictions called forecast, which includes a yhat column with the actual predicted values. (Prophet, 2023)

This report contains two types of prophet models: a baseline one with just ds and y as columns, and one with multiple lagged regressors such as the air temperature and zonal winds, which were shown to have the highest impact on the prediction of the sea surface temperature in the EDA. Both models have been assessed through the mean average error.

We will first look at the baseline model, which came with an MAE of 0.88.

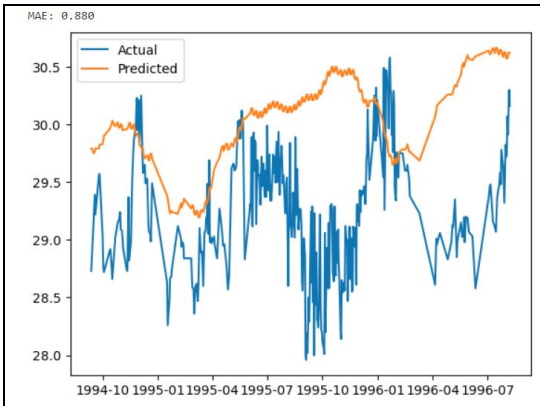
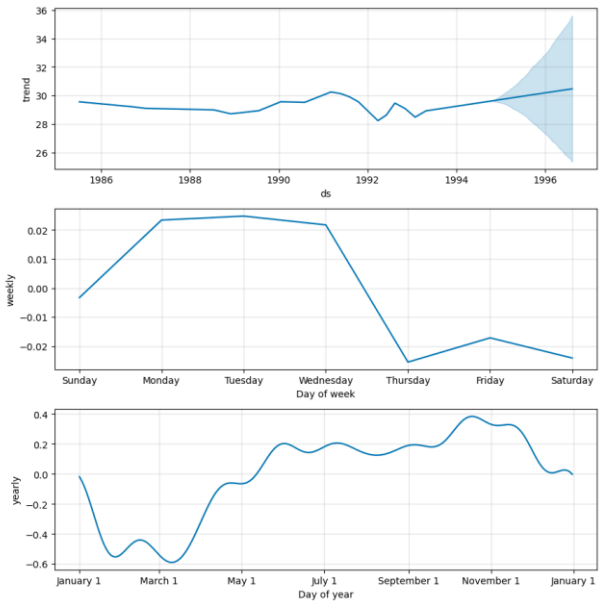


Figure 6. Prophet baseline model with just ds and y

Figure 7. Prophet baseline components



The baseline model performed surprisingly well as it tried to follow the trend in the data a bit, however it came with a large MAE which showed room for improvement, especially in the trend component. The highlighted portion in blue shows the model loses confidence when the forecast starts (after 1996).

To improve the model further we introduced two exogenous variables namely air_temp and zon_wind by creating a lagged version of them (lag window = 3). In order to be able to forecast with the model that has external regressors, we featured engineered these lagged regressors for both the training and validation sets by shifting these columns down by three days. As expected this model produced the best results. One limitation that comes with this model is that it can only forecast three steps ahead. This could be valuable for short term forecasting leads, but not very helpful if there is a need for a long term forecast.

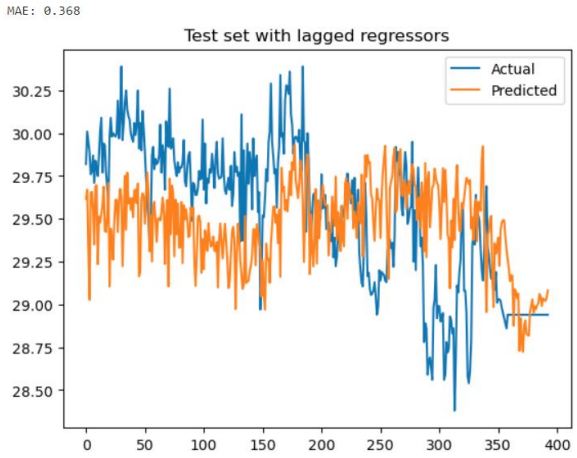


Figure 8. Prophet on test set with multiple regressors

The extra work was rewarded with an MAE of 0.368 which is significantly better than the baseline model with an MAE of 0.88. Looking at the plot above, it shows the model actually following the trend in data and overlapping significantly. As always, there is still room for improvement, such as identifying other hyperparameters to tune according to this page: *include meta page with multiple hyperparameters*.

This experiment showed the importance of feature selection in weather time series forecasts but also the ease of implementation of Meta's model Prophet, especially for data that was more complicated to collect.

2. *ARIMA*

According to (Practical Time Series Analysis, p 202), some of the advantages of statistical methods for time series are that models like ARIMA are relatively simple to use, and easy to understand in terms of parameters. They can be applied to smaller datasets and still get a good result, which was helpful in the case of this report, as we had to use a sample dataset. The authors also highlight that even though models like ARIMA are fairly simple to implement, they still prevent overfitting whilst delivering a good performance. Some of the disadvantages concern the fact that these models cannot handle nonlinear dynamics and will not be able to identify any specific trends in data – happily this was not the case for this report.

The ARIMA models were implemented using the ACF and PACF plots that highlight the parameters needed to run the models: p = 2, d = 0, q = 1.

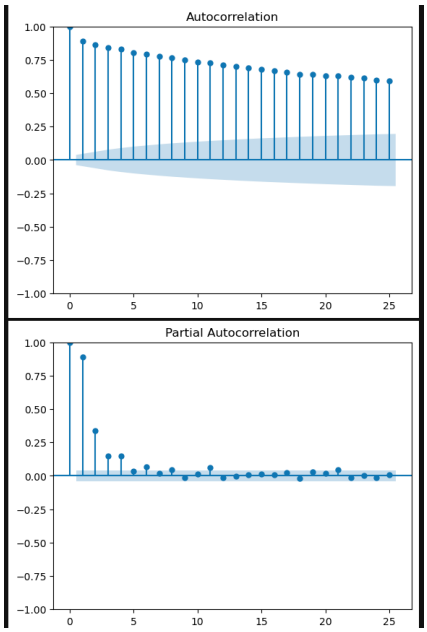


Figure 9 ACF and PACF correlation plots

According to PTSA (p 90), "autocorrelation gives you an idea of how data points at different points in time are linearly related to one another as a function of their time difference" which is why this report used the ACF plot. According to TSFP (p.193), the PACF plot (partial auto correlation function), the PACF "measures the correlation between lagged values in a time series when we remove the influence of correlated lagged values in between". According to the ACF plot, the moving average window size should be 1 and according to the PACF the lag order

of the auto-regression should be 2. However, after a grid search the best parameters on the validation set proved to be 1, 1, 2.

The baseline model for ARIMA had a MAE of 0.424. Plotting the forecast against the target indicates that the model is unable to track the seasonality of the data and thus results in what looks like forecasting the trend.

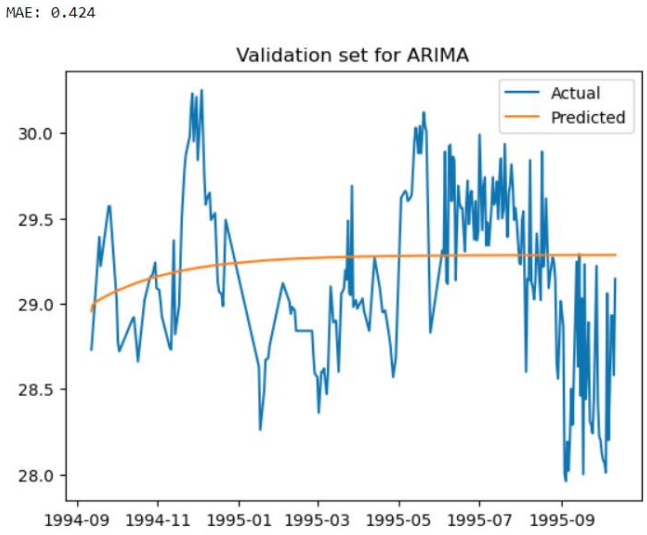


Figure 10. ARIMA done on validation set

The next step in the process was to create a grid search for ARIMA that contains the best combination of hyperparameters. This one gives out the MAE of 0.4210.

In conclusion, based on the conducted report, it seems that the Prophet model with lagged regressors has performed the best on unseen data, as it offered an MAE of 0.368, which is the lowest obtained. The expectation at the start of this report was that the Prophet model would perform best and this was indeed the case.

As for future recommendations, this report would like to focus next on using SARIMA and SARIMAX, which are models that are specifically designed for high seasonality in data.

REFERENCES:

US Department of Commerce, N.O. and A.A. (2016) *Why do scientists measure sea surface temperature?*, *NOAA's National Ocean Service*. Available at: https://oceanservice.noaa.gov/facts/sea-surface-temperature.html#:~:text=Because%20the%20ocean%20covers%2071,on%20the%20global%20climate%20system (Accessed: 25 April 2024).

Met Office (no date) *Artificial Intelligence for Numerical Weather Prediction*, *Met Office*. Available at: https://www.metoffice.gov.uk/research/approach/collaboration/artificial-intelligence-for-numerical-weather-prediction#:~:text=Machine%20Learning%20models%20are%20exceptionally,prediction%20of%20extreme%20weather%20events (Accessed: 25 April 2024).

Taccari, M.L. (2023) *Forecasting the future: Pioneering weather predictions with machine learning*, *Forecasting the Future: Pioneering Weather Predictions with Machine Learning | Leeds Institute for Data Analytics*. Available at: https://lida.leeds.ac.uk/news/forecasting-the-future/ (Accessed: 25 April 2024).

Can ai transform how we forecast the weather? (2023) *Official blog of the Met Office news team*. Available at: https://blog.metoffice.gov.uk/2023/10/31/can-ai-transform-how-we-forecast-the-weather/ (Accessed: 25 April 2024).

Newman, R. and Noy, I. (2023) *The global costs of extreme weather that are attributable to climate change*, *Nature News*. Available at: https://www.nature.com/articles/s41467-023-41888-1 (Accessed: 25 April 2024).

El Nino (no date) *UCI Machine Learning Repository*. Available at: https://archive.ics.uci.edu/dataset/122/el+nino (Accessed: 25 April 2024).

Jadon , S., Kanty, J. and Patnakar, A. (2021) *Challenges and approaches to time series forecasting: A survey*, *ResearchGate*. Available at: https://www.researchgate.net/publication/348176423_Challenges_and_Approaches_to_Time_series_forecasting_A_Survey (Accessed: 29 April 2024).

Peixeiro, M. (2022) *Time series forecasting in Python*. 1st edn. Shelter Island, New York: Manning Publications Co. LLC. , noted as “TSFP” in report

Nielsen, A. (2020) *Practical time series analysis: Prediction with statistics and machine learning*. Sebastopol, CA: O’Reilly. , noted as “PTSA” in report

Prophet Documentation (2023) *Prophet*. Available at: https://facebook.github.io/prophet/docs/quick_start.html (Accessed: 29 April 2024).