UCL DEPARTMENT OF GEOGRAPHY

YEAR 2024-25

EXAM CANDIDATE ID:	**PLXY6**
MODULE CODE:	**CEGE0042**
MODULE NAME:	**Data Mining**
COURSE PAPER TITLE:	**Report**
WORD COUNT:	**6 page-report**

Introduction

Researching sea surface temperature (SST) is crucial for monitoring climate, managing ecological systems, and understanding ocean-atmosphere interactions (Bonino et al., 2024). Forecasting SST anomalies helps improve our understanding of climate change and its impacts (Bonino et al., 2024). This is essential for managing marine biodiversity and sustaining food security in the region (Bonino et al., 2024). A recent study focuses on predicting SST and marine heatwaves using Copernicus Marine Services (CMEMS) and European Centre for Medium-Range Weather Forecasts reanalysis datasets for SST (Bonino et al., 2024). The study uses Random Forests (RF), Convolutional Neural Networks and Long Short-Term Memory models to successfully predict extreme events and anomalies (Bonino et al., 2024). This shows that combining satellite-derived SST data with machine learning (ML) improves prediction (Bonino et al., 2024). Shao et al. (2021) uses Neural Networks (NN) to predict SST in the South China Sea by applying residual error correction, a method which improves the reliability of short-term predictions (Shao et al., 2021). The integration of climate indices like Oceanic Niño Index (ONI) is analysed by Kambezidis et al. (2024). The research explores the correlation between SST and ONI in the Eastern Mediterranean and concludes that SST trends are highly affected by large-scale climate events like El Niño (Kambezidis et al., 2024). This research will build on previous studies by evaluating model performance using ML for both short and long-term SST anomaly forecasting, with a more significant focus on short-term prediction.

Methodology

This study focuses on forecasting SST anomalies for 2024 in the Adriatic Sea, which is bordered by six countries, including Italy, Greece, and Croatia (Fig. 1). The data used is from the freely accessible Copernicus Marine Services (CMEMS), called *Mediterranean Sea High Resolution and Ultra High-Resolution Sea Surface Temperature Analysis* (Fanelli et al., 2024). The dataset is sourced from satellite observations and processed by MET Norway (Fanelli et al., 2024). This high-resolution SST product has been assimilated and calibrated as it is processed at Level 4 (Fanelli et al., 2024). The data is available in daily temporal resolution and has a spatial resolution of 0.01° × 0.01° (Fanelli et al., 2024). This is essential to make the dataset reliable (Fanelli et al., 2024). Moreover, it is downloaded in NetCDF-3 format. The region is selected by drawing a polygon on CMEMS data viewer, covering the Adriatic Sea, located between the Tyrrhenian and Ionian Seas, with coordinates of 46.303°N, 11.614°E, encompassing 774 km^2.

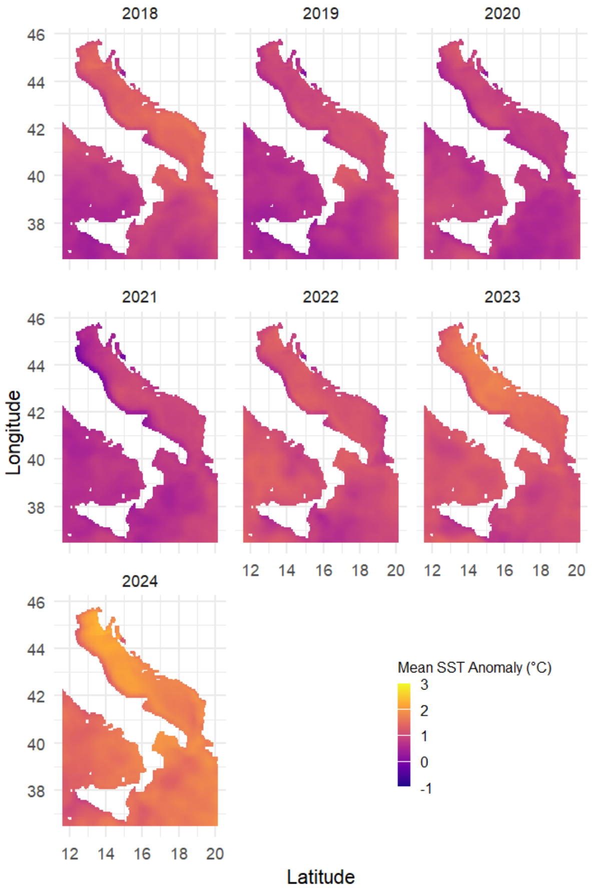

Figure 1: Heatmaps of Mean SST Anomaly in the Adriatic Sea (2018 to 2024)

The second dataset used is the climate index of Oceanic Niño Index (ONI) created by NOAA (NOAA). This index is based on the three-month running mean of SST anomalies in the Niño 3.4 region, located in the central Pacific Ocean near the equator (NOAA). It measures deviations from the long-term average SST (NOAA). Positive values indicate warm events (El Niño) whereas negatives ones represent La Niña (Fig. 2) (NOAA). The index is updated monthly and is a primary tool to analyse ENSO, crucial for long-term and short-term climate research (NOAA). The Locally Estimated Scatterplot Smoothing (LOESS) is used to create a smooth timeseries of the two datasets (Fig. 2) (Rojo et al., 2017). It fits multiple regressions to localised neighbourhoods of the data (Fig. 2) (Rojo et al., 2017). Instead of fitting a single regression to the data, LOESS creates separate regression to each section of the data (Rojo et al., 2017). This captures trends in the dataset and helps detect longer-term changes without accounting short-term variations (Fig. 2).

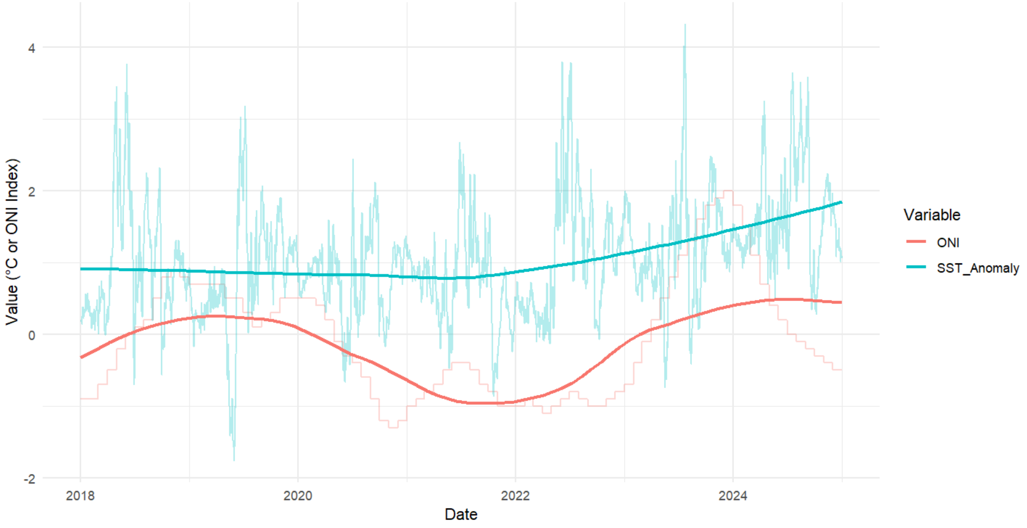

Figure 2: Daily Mean SST Anomalies (blue) and ONI (red) smoothed using LOESS (2018-2024)

These datasets are merged on RStudio to perform two ML algorithms, RF and Neural Networks (NN). RF is an ensemble ML algorithm working with multiple decision trees. During the training stage, the dataset builds the decision trees to make predictions based on random subsets of the data (Cutler et al., 2007). Then, these are combined to produce a final and more accurate prediction (Cutler et al., 2007). This method is easily interpretable and handles non-linear relationships in the dataset (Cutler et al., 2007). In this study, for both ML models, the dataset is split into training-validation, with the training set being from 2018 to 2023 and the validation set 2024 (Fig. 3). This allows for the training model to be validated using an unseen year to evaluate the model. As the aim is to find the best settings for the model, hyperparameter tuning plays an important role in RF (Cutler et al., 2007). Root Mean Square Error (RMSE) is used as the evaluation metric to find the optimal value of 'mtry' parameter.

When there is a consistent error (bias) in the model, like overestimation or underestimation, this issue is addressed using bias correction (Fig. 3) (Cannon et al., 2015). The simple bias correction corrects the predictions based on the mean difference between observed and modelled values (Fig. 3) (Cannon et al., 2015). The regression-based bias correction uses a linear regression model to estimate the differences between the predicted and observed SST values (residuals) (Fig. 3) (Cannon et al., 2015). This finds patterns in the residuals to determine a trend and adjust the model to correct them (Fig. 3) (Cannon et al., 2015).

NN are ML models which capture non-linear relationships in SST forecasting (Shao et al., 2021). It has interconnected layers of neurons which are assigned optimal weights through training and minimise error through backpropagation (Shao et al., 2021). Hyperparameter tuning modifies the decay rate and the network size to find the optimal configuration (Shao et al., 2021). The best performing NN model has a size of 5 and a decay rate of 0.001. Similar to RF, it is also adjusted using a residual correction, which improves the accuracy of the model.

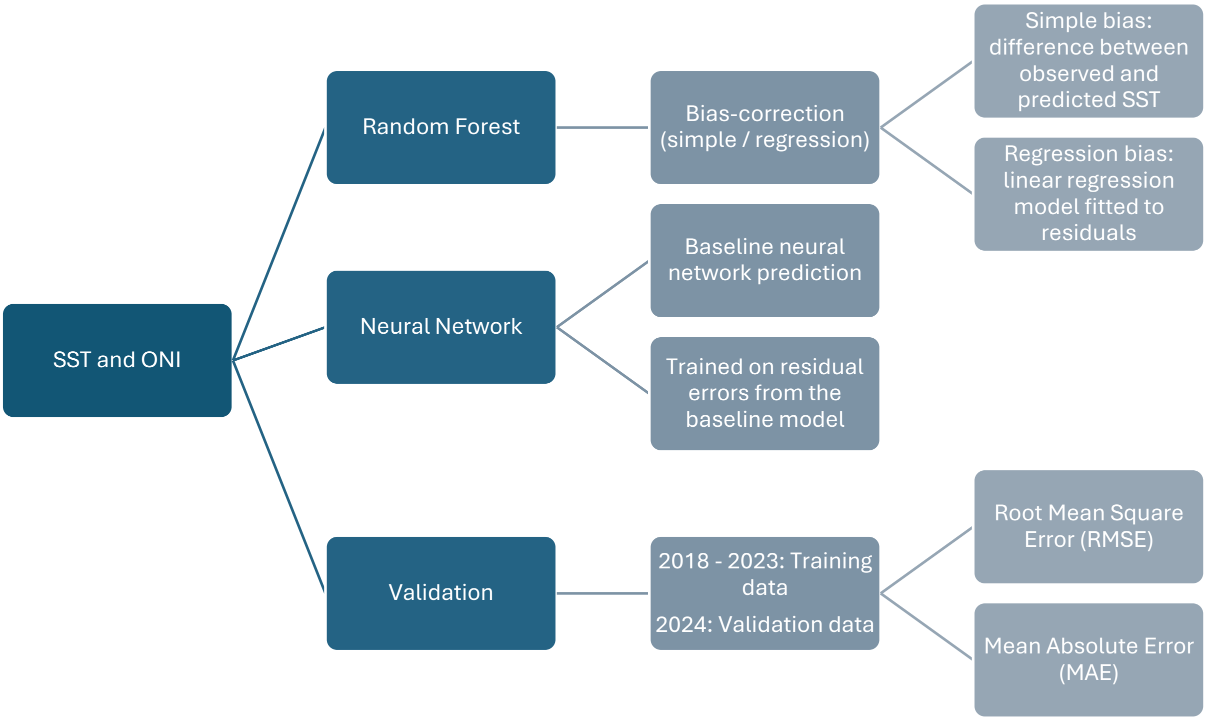
SST and ONI
Random Forest
Neural Network
Validation
Bias-correction (simple / regression)
Simple bias: difference between observed and predicted SST
Regression bias: linear regression model fitted to residuals
Baseline neural network prediction
Trained on residual errors from the baseline model
2018 - 2023: Training data
2024: Validation data
Root Mean Square Error (RMSE)
Mean Absolute Error (MAE)

Figure 3: Workflow for forecasting sea surface temperature (SST) in the Adriatic Sea using machine learning models

Results

Table 1: RF and NN Models' MAE and RMSE results

Model/Approach	**MAE**	**RMSE**
RF + Simple Bias	0.74	0.99
RF + Regression-Bias Correction	**0.52**	**0.64**
NN Baseline	0.60	0.76
NN Residual Correction	0.60	0.76

The model performance is evaluated using RMSE and Mean Absolute Error (MAE). On the one hand, for RF models, the regression-based bias correction performs best, with an RMSE of 0.64 and a MAE of 0.52 (Table 1). For the simple bias model, the RMSE is higher by 0.35 and the MAE higher by 0.21 (Table 1). On the other hand, for NN, the difference is not as large. The residual model has an RMSE of 0.76 and a MAE of 0.60, which are lower by 0.001 for the RMSE and 0.003 for the MAE compared to the baseline model (Table 1). The parameters used for both models are from the best performing NN model, with a size of 5 and a decay rate of 0.001.

RF and NN can be compared based on the residual models as they show better results than baseline ones (Table 1). In the first month, RF performs better, but declines after March (Fig. 4). In contrast, NN perform better later in the year, specifically during summer months (Fig. 4).

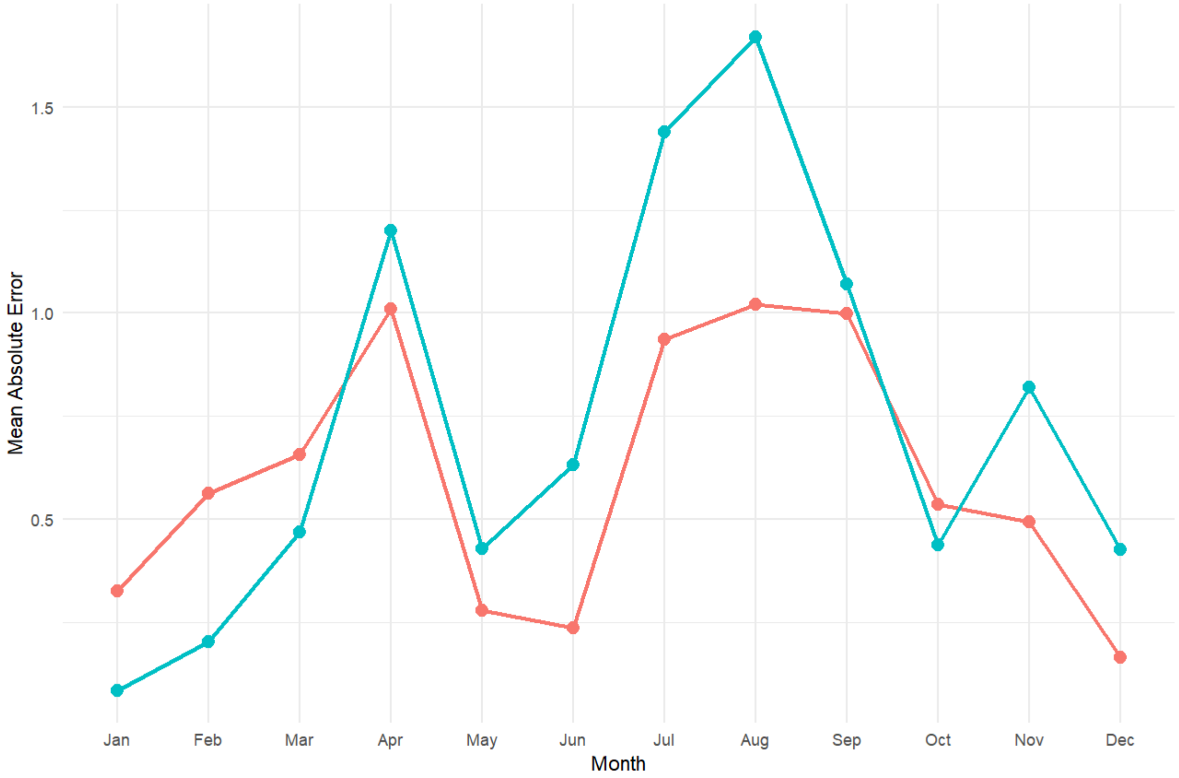

Figure 4: Monthly Mean Absolute Error. Red: NN, Blue: RF.

NN has a strong short-range memory, also seen through the Autocorrelation Function (ACF) (Fig. 4 and 5). This is very high for the first lags before a gradual decrease (Fig. 5). This pattern shows SST anomalies on one day are likely to remain similar on subsequent days, which reflects the persistence of the pattern over short periods (Fig. 5). When the ACF values cross the threshold for no correlation (zero correlation - horizontal lines), the correlation is statistically significant (Fig. 5). This means the pattern is accurate, not based on random variations (Fig. 5).

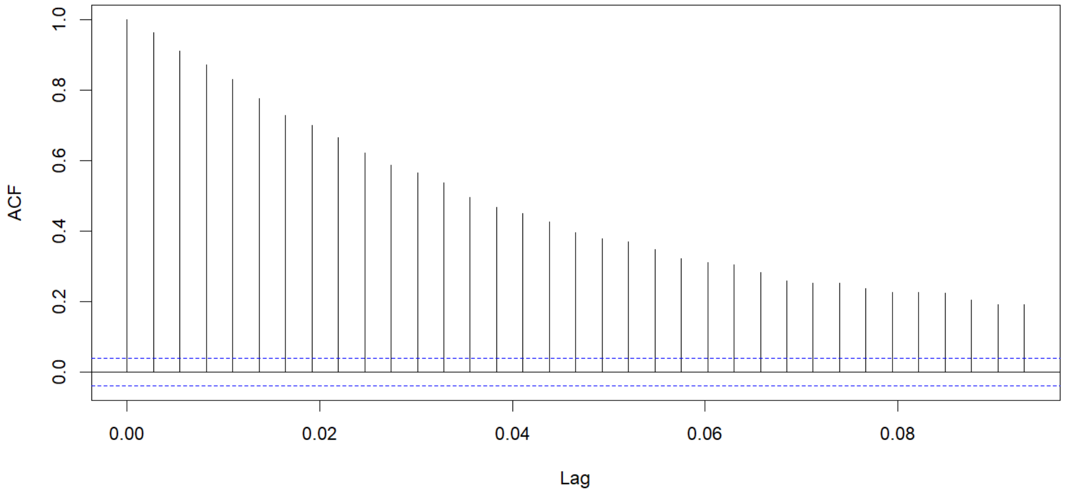

Figure 5: Autocorrelation plot. Horizontal lines: zero correlation.

To further evaluate the models' performance, baseline predictions and observed SST anomalies can be compared for January 2024 (Fig. 6 and 7). Both models' bias correction (in red) shows differences between the predicted values (x-axis) and the observed ones (y-axis) (Fig. 6 and 7). These systematic errors show predictions over or underestimate the SST anomalies (Fig. 6 and 7). Nevertheless, applying bias correction (in red) makes a significant improvement in the model's accuracy by adjusting the model's predictions (Fig. 6 and 7). This reduces the gap between observed and predicted values, fitting the data with a linear relationship for short-term predictions. The NN model has more variability and deviation from observed values, especially at lower SST anomalies, meaning it underestimates the values (Fig. 6 and 7). The RF model has a more consistent trend, showing a higher correlation between values (Fig. 6 and 7).

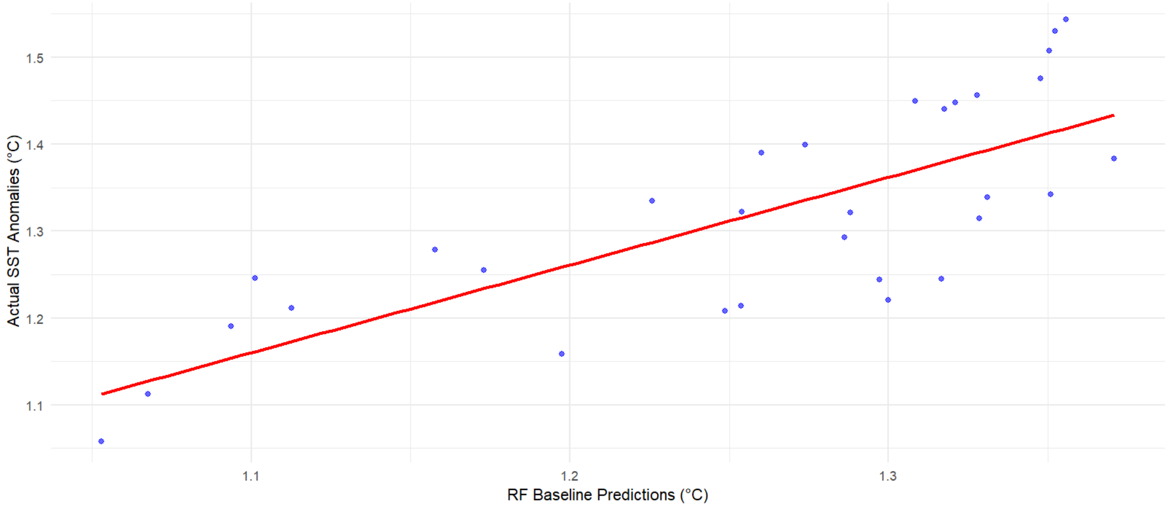

Figure 6: RF scatter plot - Predictions vs Observed SST Anomalies (January 2024). Blue: predicted values, Red: Bias correction fit.

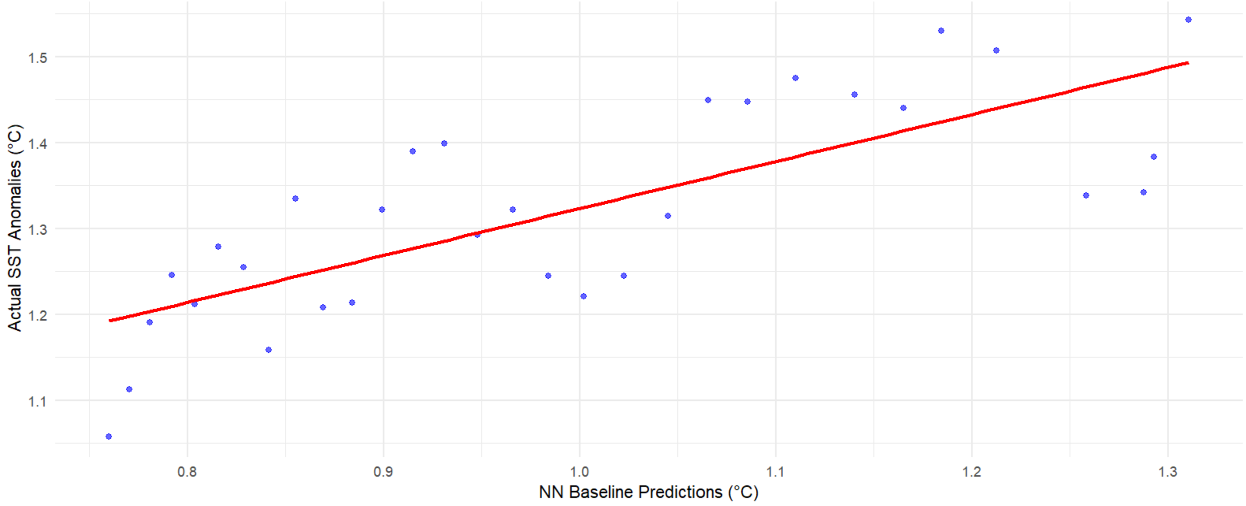

Figure 7: NN scatter plot - Predictions vs Observed SST Anomalies (January 2024). Blue: predicted values, Red: Bias correction fit.

Discussion and Conclusion

The SST data originating from CMEMS was chosen to assess both short-term variabilility and long-term trends. Copernicus provides high temporal and spatial resolution which is crucial for capturing fine-scale variability (Fig. 1). In a small-scale region like the Adriatic and Ionian Seas, the temperature anomalies are localised, and the chosen dataset can spot this (Fig. 1). The narrow shape of the Adriatic Sea and its proximity to six different countries' lands make the region an engaging study (Fig. 1). The addition of the climate index ONI provides a crucial context for SST anomalies, providing broader climate factors such as El Niño and La Niña events to the models (Fig. 2).

Looking at the temporal aspect, the dataset chosen is from 2018 to 2024, split for training and validation (Fig. 3). This is useful for model evaluation and for prediction. The dataset is updated twice a day (12:00 and 15:00 UTC) which creates accurate short-term predictions. Nevertheless, when long-term forecasting is analysed, five years of data, chosen due to computational constraints, is insufficient to create a reliable model.

The overall performance of the models finds that RF is more reliable until Mid-March, but displays high MAE from then until mid-September (Fig. 4). NN models exhibit lower MAE during summer months and from mid-October to December making the model more accurate to depict seasonality (Fig. 4). Looking at statistical results, RF has a lower RMSE and MAE, indicating it has a higher overall performance for long-term compared to NN (Table 1). This is also consistent with previous studies in using non-linear data (Cutler et al., 2007).

Focusing more on methods, the bias correction technique (regression-based correction) improved the RF model's accuracy by correcting systematic prediction errors due to unreliable trends in the residuals (Table 1). In contrast, the NN residual correction only displays minimal improvements, showing that NN did not have biases in its predictions (Table 1).

From a short-term prediction perspective, the autocorrelation plot is essential to understand the models' predictability (Fig. 5). SST anomalies are very similar for the first days, as shown with the high lags, and are significantly correlated with previous values (Fig. 5). Therefore, forecasting models rely on subsequent days for prediction (Fig. 5). The use of lag features also plays a key role in recording temporal correlations in the data, crucial for short-term forecasts (Fig. 5).

This analysis shows the necessity of ML models to refine continuously for tasks such as SST forecasting. The choice of the model depends on short or long-term forecasting, and in this research focusing on January, RF is more reliable (Fig. 6 and 7).

References

Bonino, G., Galimberti, G., Masina, S., McAdam, R. and Clementi, E., 2024. Machine learning methods to predict sea surface temperature and marine heatwave occurrence: a case study of the Mediterranean Sea. *Ocean Science*, 20(2), pp.417–432. Available at: https://doi.org/10.5194/os-20-417-2024.

Cannon, A.J., Sobie, S.R. and Murdock, T.Q., 2015. Bias correction of GCM precipitation by quantile mapping: How well do methods preserve changes in quantiles and extremes? Journal of Climate, [online] 28(18), pp.6938–6959. Available at: https://doi.org/10.1175/JCLI-D-14-00754.1.

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. and Lawler, J.J., 2007. Random forests for classification in ecology. Ecology, [online] 88(11), pp.2783–2792. Available at: https://doi.org/10.1890/07-0539.1.

Fanelli, C., Ciani, D., Pisano, A. and Buongiorno Nardelli, B., 2024. Deep learning for super-resolution of Mediterranean sea surface temperature fields. EGUsphere, [pre-print] 2024, pp.1-18. Available at: https://doi.org/10.5194/egusphere-2024-1234.

Kambezidis, H.D., 2024. Atmospheric Processes over the Broader Mediterranean Region: Effect of the El Niño–Southern Oscillation? Atmosphere, [online] 15(3), p.268. Available at: https://doi.org/10.3390/atmos15030268.

NOAA Climate Prediction Center, 2025. Oceanic Niño Index (ONI). [online] Available at: https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php [Accessed 1 March 2025].

Rojo, J., Rivero, R., Romero-Morte, J. et al., 2017. Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing. International Journal of Biometeorology, 61(3), pp.335–348. Available at: https://doi.org/10.1007/s00484-016-1215-y.

Shao, Q., Li, W., Han, G., Hou, G., Liu, S., Gong, Y. and Qu, P., 2021. A Deep Learning Model for Forecasting Sea Surface Height Anomalies and Temperatures in the South China Sea.

Journal of Geophysical Research: Oceans, [online] 126(6). Available at:

https://doi.org/10.1029/2021JC017515 [Accessed 27 March 2025].

Appendix

Data Source and Access Instructions

The sea surface temperature anomaly data used in this project is obtained from the Copernicus Marine Environment Monitoring Service (CMEMS). Specifically, the datasets for the Adriatic Sea were accessed via the Copernicus data viewer, which provides free access to the data in netCDF format. You can download the data directly by visiting the following link:

https://data.marine.copernicus.eu/-/tcxi92iyo2

On the data viewer page, you will find a download button that allows you to select and download the required files. In my project, I used three datasets corresponding to different time periods (2018–2020, 2021–2023, and 2024). The filenames used in the code (e.g., AdriaticSea_2018-2020.nc) correspond to the downloaded files. For reproducibility, the code includes instructions on how to set the working directory and load these files. You can follow the above link and download the data manually.