Analysis of Sea Surface Temperature north of the Canary Islands and modeling of an out-of-normal event with reanalysis data contrasted with in-situ data

Introduction to Data Analysis and Machine Learning

Sergio Sicilia González[a]

[a] *UIB-IFISC, Mallorca, Spain*

Abstract

This paper explores the study of Sea Surface Temperature (SST), a key parameter for understanding global climate dynamics, weather patterns and marine ecosystems. In this study, SST is examined as a time series, taking advantage of reanalysis data from the Copernicus Marine Service. This dataset is compared with in situ observations obtained from Argo buoys to ensure reliability. A historical analysis of these time series is carried out, highlighting a distinct episode of unusually high SST that occurred over several months in 2023. To encapsulate the complexity of these phenomena, we employ advanced modeling techniques, including autoregressive time series models and recurrent neural networks. These methods are applied not only to represent the single high temperature event of 2023, but also to capture the broader shades of the SST time series. The resulting models are evaluated against direct measurements from Argo buoys, providing robust validation of our analytical approach and contributing to a deeper understanding of SST variations and predictive models. [1]

Keywords: Sea Surface Temperature, Time Series, Recurrent Neural Networks, Argo Floats, Reanalysis

Contents

1 Introduction **2**

2 Methodology and Data **2**
2.1 Data . . . 2
2.1.1 Argo Floats . . . 2
2.1.2 Copernicus Marine Service . . . 3
2.2 Preparing the data . . . 3
2.3 Forecasting tools . . . 4
2.3.1 SARIMAX . . . 4
2.3.2 LSTM . . . 4

3 Results and Discussion **5**
3.1 SST series . . . 5
3.2 SARIMAX forecast . . . 6
3.3 LSTM forecast . . . 7

4 Conclusions **9**

1. Introduction

This study has been motivated by recent events in the North Atlantic Ocean during the summer of 2023, where records of statistical deviation (anomaly) have been broken. The causes are not fully known and are currently under investigation, but it is believed to be a multiple cause, among others to a weakening of the Azores, with values between 20 and 30% lower than usual at the time, allowing a rapid increase in Sea Surface Temperature and affecting the Canary Current. Also due to the reduction of Saharan dust in the north, thus reducing the vertical mixing and allowing the increase of temperatures. Moreover due to the large scale effects such as global warming. As an effect of these temperatures the hurricane season have been boosted, which during the month of June produced 3 hurricanes, something exceptionally rare for the time. These events have been of a very short duration compared to the El Niño oscillation (which is just beginning). Together with its initial effects, it is predicted higher than average temperatures for the first months of 2024 [2].

A major problem underlying climate change is the complication of (small-scale) weather forecasting, which is why new techniques must be developed to predict and understand these new and unusual events. Several time series prediction techniques are presented, one is an autoregressive integrated moving average model (ARIMA) and the other is a more advanced algorithm of recurrent neural networks called Long Short-Term Memory. In addition, a historical analysis of the time series in the north of the Canary archipelago is made, with the aim of understanding the oddity of this phenomenon occurred in the last months.

These algorithms have been tested in different articles, as [3]. Where the researchers propose a prediction method using sea surface temperatures (SSTs) and deep-learning technology, employing a Long Short-Term Memory (LSTM). The proposed model's performance, assessed against existing SST prediction models and external meteorological data, showed superior accuracy in predicting SSTs, although with reduced effectiveness over longer prediction intervals. Or in [4], where the researchers did the same but for several others Machine Learning algorithms. The findings reveal that the integration of automated feature engineering with machine-learning approaches yields accuracy on par with current advanced models. The models successfully capture seasonal trends and accurately reflect short-term variations influenced by atmospheric conditions. This study highlights the efficacy of machine learning-based methods as versatile prediction tools for ocean variables.

2. Methodology and Data

In this section we explain the data used, its collection and processing, as well as the procedures carried out in the analyses.

2.1 Data

The two sources of data for this work are presented. As well as their characteristics and data collection process.

2.1.1. Argo Floats

Argo is an international program that collects information from the interior of the ocean using a fleet of robotic instruments that drift with ocean currents and rise and fall between the surface and an average water level. Each Argo float is launched from a ship. The weight of the float is carefully adjusted so that, as it sinks, it eventually stabilizes at a preset level, typically 1 km. Ten days later, an internal pump driven by a battery transfers oil between a reservoir inside the float and an external bladder. This causes the float to first descend to 2 km and then return to the surface, measuring ocean properties as it rises. The data and the float's position are transmitted to satellites and then to

receiving stations on land. The float then sinks again to repeat the 10-day cycle until its batteries are depleted.

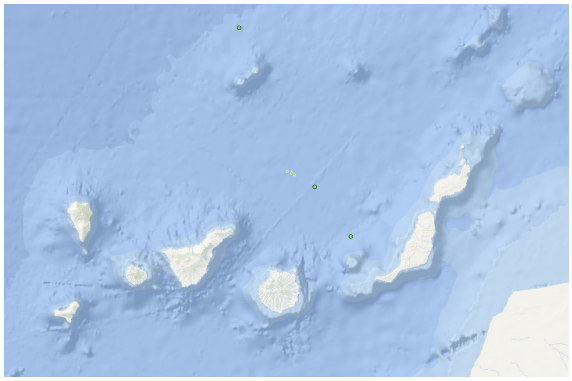

Figure 1: The green circles are the current position (at the end of December 2023) of the Argo floats in the Canary archipelago.

The floating data we are going to collect is located at [5]. Figure 1 shows the 3 buoys that currently exist in the Canary Islands. The one used for the analysis is in the upper part of the figure. These floats have been placed by the research staff of the Spanish Institute of Oceanography based in Santa Cruz de Tenerife. The one used in the study was launched on April 24, 2023 from the *R/V Ángeles Alvariño* during the *RAPROCAN2304* cruise, as well as the other two present in the image.

2.1.2. Copernicus Marine Service

The daily numerical outputs of the SST from 1993 to December 22, 2023 were taken from Marine Copernicus Atlantic Ocean Analysis and Forecasting - Iberian Biscay Irish (IBI) [6]. Nologin runs the model on a daily basis in collaboration with Puertos del Estado and the Centro de Supercomputación de Galicia (CESGA) provides the computing resources. The dataset provides a 5-day hydrodynamic forecast that includes significant high frequency processes that are essential to characterize marine operations in the region (e.g. tidal forces, high frequency wave and atmospheric forces, river runoff). In addition, a weekly update of the truncated IBI analysis is available as an enhanced historical IBI assessment (reanalysis). The system is based on the eddy-resolving NEMO model operating at a horizontal resolution of $1/12^\circ$ ($\approx 9-10$ km).

2.2 Preparing the data

Before doing any analysis, adjustments had to be made to the data, as there was a lot of information (related to the study of the ocean) that was not of interest to us. To begin with, the databases were in *NetCDF* format [7], commonly used for data related to atmospheric, oceanic and general environmental variables.

With respect to the buoy data, the measured temperatures were read and only those at sea surface level were selected. The trajectory followed by the buoy was also studied, as well as the stations where measurements were taken. This can be seen in Figure 2, where the average distance between stations is 27 ± 13 km. Since, unlike the Copernicus data, the buoy data is not synoptic, i.e., they are not constant in space, we then approximate the locations of the different stations to the red circle in the figure. It can be seen in the color map in Figure 2 that the difference in mean temperature during this period between the initial and final locations is approximately 0.5 ºC, which is a fair price to pay in exchange for substantially simplifying the model.

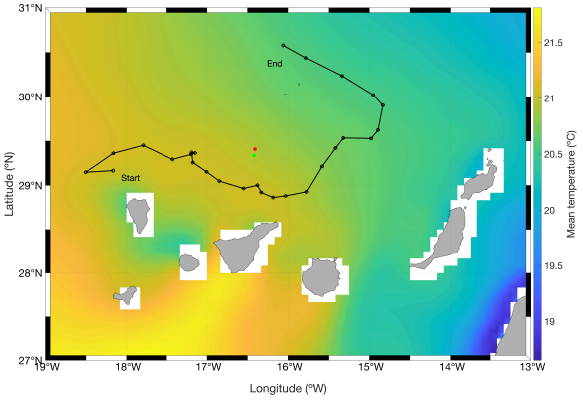

Figure 2: Buoy track from April to December. The black circles are the sampling sites, the red circle is the mean location of those stations and the green is the nearest node of the Copernicus reanalysis grid. The color map is the mean SST (calculated using the Copernicus data) during the months of the buoy path.

In addition, since we want to make a comparison with the reanalysis data, we will use the SST data from *Copernicus* located in the green circle, since it is the closest to the red one, being at about 5 km, which is lower than the model resolution, and therefore a good approximation.

Another drawback with respect to the Copernicus data was that at the time of the analysis, the data was not available in the same database. Since in one of them there was only information from 1993 to 2021 and another one from 2018 to the end of 2023, so both were used, the information from the green node was selected and the common data was filtered, to finally have a single time series from 1993 to 2023 in the desired location.

2.3 Forecasting tools

The two algorithms used for SST prediction are presented here. Because it is outside the scope of the paper, the explanation will not be rigorous nor detailed, especially section *2.3.2*, since it is not one of the algorithms that we have properly seen in the course.

In addition, for the model explained in *2.3.1*, a reduction of data per week was made, that is, instead of having 365 annual data, there will be approximately 52. This is due to the fact that the algorithm is not prepared for so much data, causing the training time to be extremely long, or directly leading to the death of the jupyter notebook kernel. In the next section we will discuss the validity of this approach.

Temperature data from 1993 to 2020 was used to train the models, and predicted from 2020 to 2023. The latter year being our SST anomaly episode.

2.3.1. SARIMAX

SARIMAX [8] (Seasonal Auto-Regressive Integrated Moving Average with eXogenous regressors) is a statistical model widely used for the prediction of time series with seasonality. This model consists of three components. The autoregressive (AR) element relates the current value to past values (lags). The moving average (MA) element assumes that the prediction error is a linear combination of past prediction errors. Finally, the integrated component (I) indicates that the values of the original series have been replaced by the difference between consecutive values.

SARIMAX models extend the ARIMA framework by incorporating seasonal patterns and exogenous variables.

In the notation of the ARIMA-SARIMAX model, the parameters p, d, and q represent the autoregressive, differencing, and moving average components, respectively. P, D, and Q are the same components for the seasonal part of the model and m the number of periods in each season.

When these terms are zero and no exogenous variables are included, the SARIMAX model is equivalent to an ARIMA.

2.3.2. LSTM

The neural network model that has been built for our purpose is a Recurrent Neural Network (RNN) specifically using Long Short-Term Memory (LSTM) units [9]. This model is used to predict future values in a time series. The components of the model are as follows:

Recurrent Neural Network (RNN): RNNs are a class of neural networks designed to recognize patterns in sequences of data, such as time series, text sequences, audio recordings, etc. They have the unique ability to retain past information through internal loops in the network. This makes them suitable for tasks where temporal context is important.

Long Short-Term Memory (LSTM): LSTMs are an extension of RNNs that solve the problem of long-term dependency (i.e., the ability to remember information for long periods of time). LSTM units have a more complex structure than standard RNN neurons, with three "gates" (the forgetting gate, the input gate, and the output gate) that

regulate the flow of information. In the context of time series, LSTM networks can learn to recognize important patterns in the training data and predict future values of the series. They are particularly well suited for predicting time series with complex patterns and long-term dependencies.

The architecture of the Model is as follows:

- LSTM Layer: The LSTM layer has 4 units (or neurons). Each LSTM unit can learn and remember information over time sequences, which is crucial for time series prediction.
- Dense Layer: After the LSTM layer, there is a dense (or fully connected) layer that reduces the output of the LSTM layer to the desired dimension. In this case, it has only 1 neuron since we are doing univariate prediction (predicting a single continuous value of the time series).

The Training and Prediction process is the following:

- Training: during training, the model adjusts its internal weights to minimize the difference between the predictions and the actual values (the loss function). We use the optimization algorithm 'adam' and the mean square error loss function.
- Prediction: Once trained, the model can make predictions on new data. The idea is to feed the model with a sequence of temperature values and allow it to predict the next value in the series, with a desired prediction time. In our case, it has been provided with a 7-day sequence of temperature and asked to predict 14 days ahead.

3. Results and Discussion

First we will make a description of the time series obtained with the Copernicus data, in order to understand the anomalous episode of this 2023, as well as basic considerations of the series.

3.1 SST series

Figure 3 shows this SST value for the entire historical series.

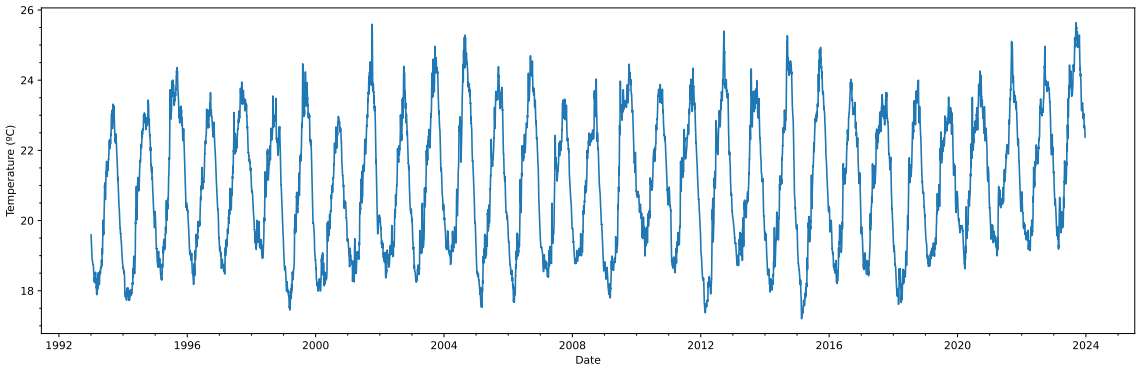

Figure 3: Temporal series of the SST, calculated using Copernicus Reanalysis, in the north of the Canary Islands.

The values in Figure 3 have been compared with those values obtained by the Argo buoy, for the days in which there are measurements, obtaining as a result a root mean square error (RMSE) equal to 0.13 ºC, a reasonable value for the approximations we have made and taking into account that the values of the Copernicus series are reanalysis and therefore not direct experimental measurements.

This time series has the characteristic

that it can be considered a Brownian signal, as can be seen in the slope of the fit in Figure 4 . This already gives us an idea of how it will behave, as well as its possible values for the SARIMAX model.

Also, the stationarity of the series has been studied by means of the Augmented Dickey-Fuller test, resulting in $p = 1.8 \cdot 10^{-25}$, and therefore it can be considered constant.

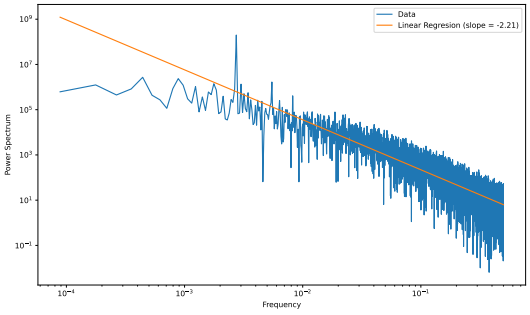

Figure 4: Power Spectrum of the 3 series.

The autocorrelation of the series tells us how this signal behaves with itself with a specific time delay. Figure 5 reveals that the SST has a high linear correlation for short lags, something that makes sense in an ocean variable where there are generally no explosive changes in temperature at short times. And for high lag times, it behaves in a periodic manner with a periodicity of 365 days, which indicates the existence of seasons, something that is also well known.

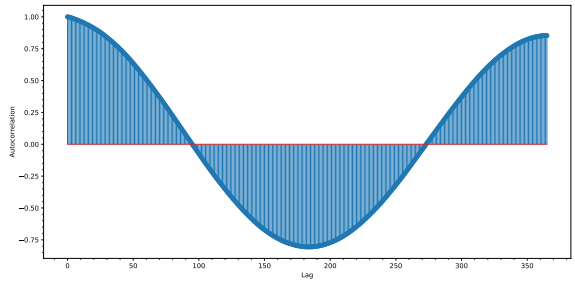

Figure 5: Autocorrelation function of the 3 series.

To understand this increasing trend in recent years, we can look at Figure 6, which shows the SST anomaly (deviation from the mean), where it can be seen that the years 2022 and 2023 are characterized by a positive anomaly practically throughout the whole year.

Furthermore, to understand the aforementioned episode during 2023, we can look at Figure 9, which shows how the temperature exceeds, on multiple occasions, 2σ of the SST for its corresponding time of the year. Especially in July, where it reaches a value of 3σ.

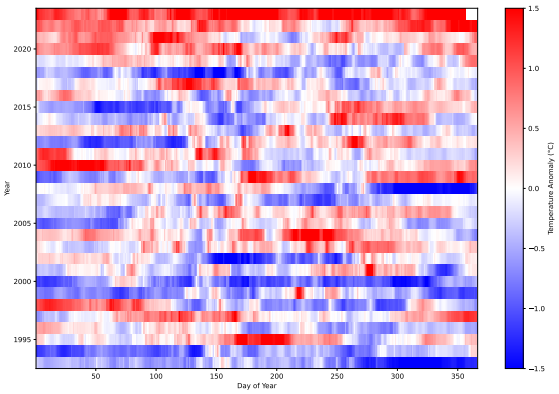

Figure 6: SST anomaly for temperatures averaged from 1993-2023.

3.2 *SARIMAX forecast*

As mentioned in the previous section, using all the daily data was unfeasible for this type of algorithm, so a weekly average was made, thus obtaining a periodic signal with value $m = 52$ in the Sarimax model. This approximation can be seen in Figure 7, where in addition, calculating the linear interpolation of the averaged data to the daily data we obtain a $RMSE = 0.017$ ºC so we can assume that this approximation is sufficiently representative of the original data.

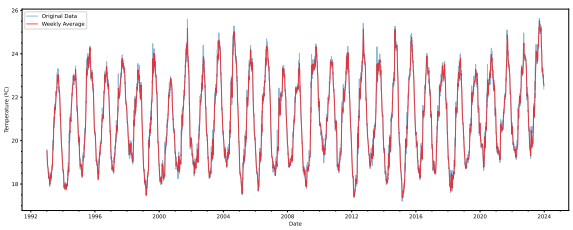

Figure 7: In blue the original Copernicus data, in red the weakly approximation.

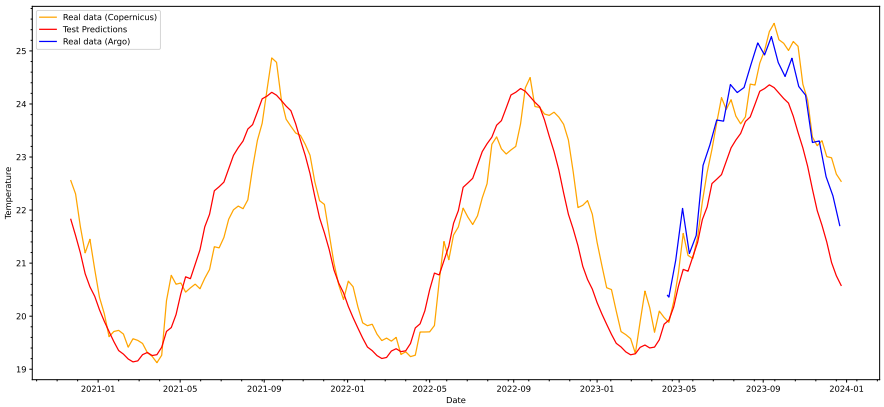

Figure 8: The red line represents the prediction on the testing set by the Sarimax algorithm for the Copernicus data, the orange line the weekly averaged Copernicus data and the blue line the values obtained by Argo.

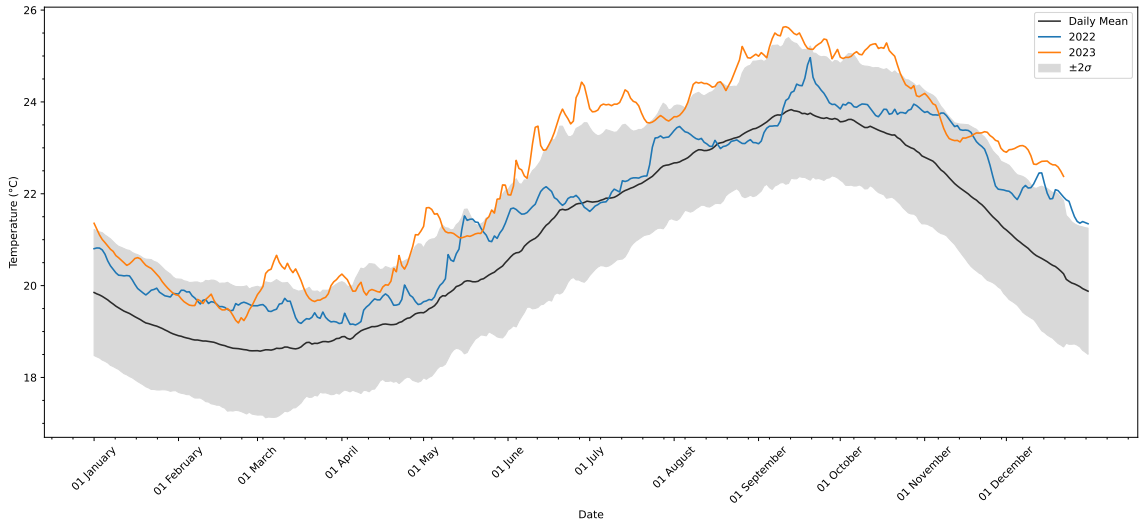

Figure 9: The black line is the daily averaged time series from 1993-2023, the gray bands its standard deviation, the orange and blue lines are the years 2023 and 2022 respectively.

The obtained optimal values of the SARIMAX parameters are all equal to 1, except the already mentioned value of stationarity m and the value of moving average ($q = 2$), these optimal values have been obtained by a combination of trial and error, and by performing tests such as those of partial autocorrelation.

In Figure 8 we can observe this prediction, which in general behaves well for those periods within the ordinary, however, it is not able to foresee an unexpected event such as those of 2023. The RMSE value between the predictions interpolated to the dates of the Argo data is equal to 1.1 ^{o}C, which indicates that it is not a correct predictor for this type of events. On the other hand, the RMSE between the testing data and its prediction is 0.56 ^{o}C, which is higher than usual due to this episode.

3.3 LSTM forecast

Due to the simplicity of this model with respect to the architecture and the type and amount of data, the loss function quickly reaches a local extreme, as can be seen in Figure 10 for the number of Epochs (learning iterations and readjustment of the weights). Since this algorithm is not the main analysis of the work, there has not been a great effort to optimize the network.

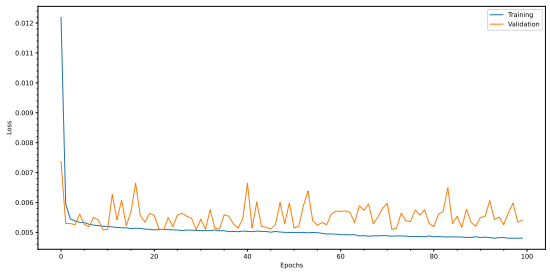

Figure 10: Number of Epochs for the Training and Validation (test) set.

Other optimizers have also been tested, as well as different batch sizes (1 in our case) and with the same number of Epochs the same end is reached.

The result for both the training and testing set prediction can be seen in Figure 11 and Figure 12, respectively. The RMSE value of the 14-day prediction with the testing set is 0.37 ºC and that of the Argo data 0.42 ºC. A much more accurate prediction than with SARIMAX.

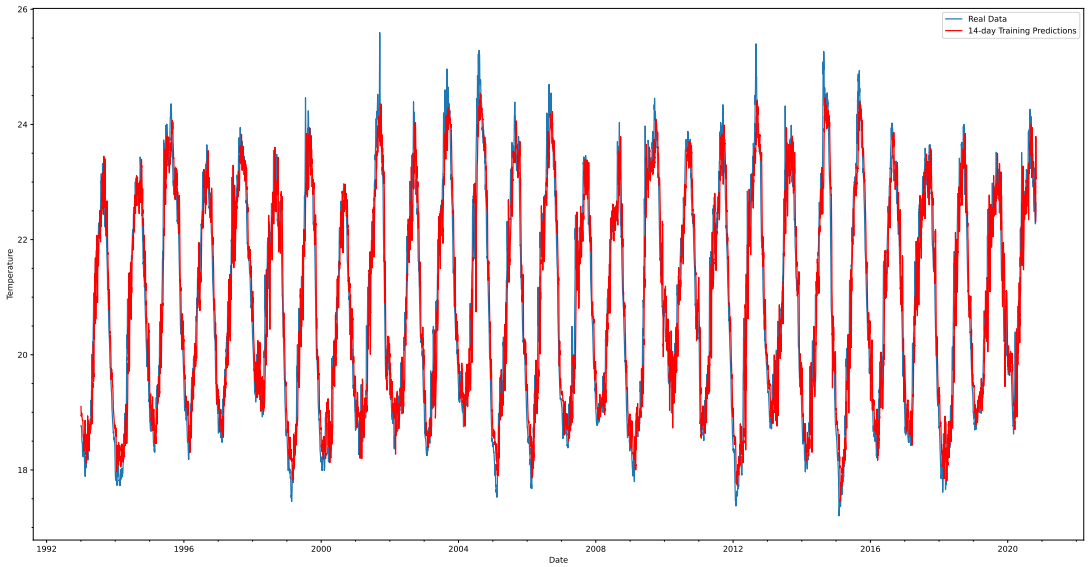

Figure 11: The blue line is the Copernicus data from training set and the red line is their prediction.

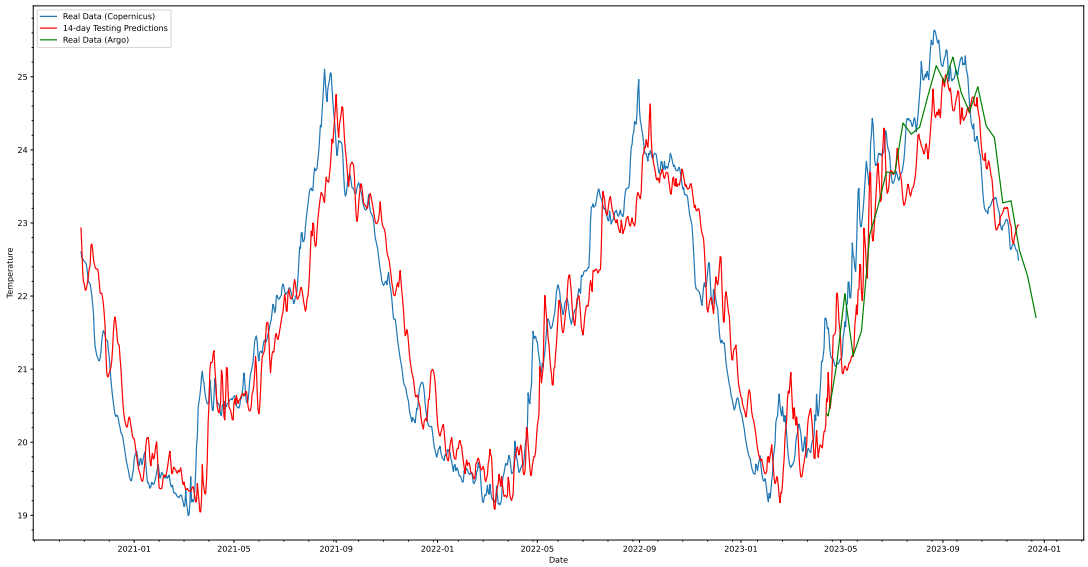

Figure 12: The blue line is the Copernicus data from testing set, the red line is its prediction and the green line is the Argo data.

4. Conclusions

We have confirmed the presence of anomalous Sea Surface Temperature behavior over much of the northern hemisphere Atlantic Ocean, a phenomenon that is not unique to this region. The causes, although partially known, require a deeper understanding. This paper highlights the complexity and unpredictability of such events, as evidenced in the Mediterranean, which has not been spared from this occurrence, where during July 2023 temperatures up to approximately 6°C above average were recorded in areas such as the coasts of Italy, Greece and North Africa [10]. This pattern suggests a broader climate change, possibly influenced by both anthropogenic and natural factors.

Historical analysis of SST reveals well-known but interesting features, such as the Brownian nature of SST, which in many instances can be considered constant and with a certain periodicity.

Forecasting models offer mixed insights. On one hand, ARIMA models show limitations when dealing with outlier events, because these values fall outside the necessary stationarity. On the other hand, Recurrent Neural Networks have shown higher efficiency, although their RMSE of 0.42 ^{o}C still does not reach the accuracy of the reanalysis model approximation ($RMSE = 0.13$ ^{o}C). This result highlights the opportunity to improve the RNNs, possibly through the integration of more extreme event data and optimization of network parameters. The scarcity of extreme event data such as that observed in the summer of 2023 highlights the importance of more robust data collection to train more accurate predictive models.

This study not only provides a detailed understanding of current STT patterns in the northern Canary archipelago, but also raises important questions about climate variability and its prediction. The need for more sophisticated models and a better understanding of oceanic and atmospheric dynamics becomes evident in these changing times. Future research should focus on the integration of more varieties of data, including those of extreme events, and on the development of more advanced predictive models that can adapt to the changing nature of marine climate.

References

[1] Sergio Sicilia Glez. Github tsa-sst. https://github.com/MisterMito/TSA-SST.git.

[2] Copernicus Climate Change Service. Record-breaking north atlantic ocean temperatures contribute to extreme marine heatwaves - news (date: 6th july 2023). https://climate.copernicus.eu/record-breaking-north-atlantic-ocean-temperatures-contribute-extreme-marine-heatwaves.

[3] Minkyu Kim, Hyun Yang, and Jonghwa Kim. Sea surface temperature and high water temperature occurrence prediction using a long short-term memory model. *Remote Sensing*, 12(21), 2020.

[4] Stefan Wolff, Fearghal O'Donncha, and Bei Chen. Statistical and machine learning ensemble modelling to forecast sea surface temperature. *Journal of Marine Systems*, 208:103347, 2020.

[5] Argo floats. Current location and buoy data. https://www.ocean-ops.org/board?t=argo#.

[6] Copernicus Marine Service Data. Atlantic-iberian biscay irish- ocean physics reanalysis. https://data.marine.copernicus.eu/product/IBI_MULTIYEAR_PHY_005_002/description.

[7] University Corporation for Atmospheric Research. Network common data form (netcdf). https://www.unidata.ucar.edu/software/netcdf/.

[8] TSA references. Sarimax. https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html.

[9] Keras. Lstm layer. https://keras.io/api/layers/recurrent_layers/lstm/.

[10] Copernicus Climate Change Service. Record temperatures in the mediterranean sea in july. https://www.copernicus.eu/en/media/image-day-gallery/record-temperatures-mediterranean-sea-july.