Machine learning methods to predict Sea Surface Temperature and Marine Heatwave occurrence: a case study of the Mediterranean Sea

Giulia Bonino[1,*], Giuliano Galimberti[2], Simona Masina[1], Ronan McAdam[1], and Emanuela Clementi[1]

[1]CMCC Foundation - Euro-Mediterranean Center on Climate Change, Italy.
[2]Department of Statistical Sciences, University of Bologna, Bologna, Italy
[*]giulia.bonino@cmcc.it

Abstract. Marine heatwaves (MHWs) have significant social and ecological impacts, necessitating the prediction of these extreme events to prevent and mitigate their negative consequences and provide valuable information to decision-makers about MHW-related risks. In this study, machine learning (ML) techniques are applied to predict Sea Surface Temperature (SST) time series and Marine Heatwaves ~~(MHWs)~~ in 16 regions of the Mediterranean Sea. ML algorithms, including Random Forest (RForest), Long short-term memory (LSTM), and Convolutional Neural Network (CNN), are used to create competitive predictive tools for SST. The ML models are designed to forecast SST and MHWs up to 7 days ahead. For each area, we performed 15 different experiments for ML techniques, modifying the training and testing periods. Alongside SST, other relevant atmospheric variables are utilized as potential predictors of MHWs. Datasets from the European Space Agency Climate Change Initiative (ESA CCI SST) v2.1 and the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis from 1981 to 2021 are used to train and test the ML techniques. ~~The~~ For each area the results show that ~~ML methods, particularly RForest and LSTM, performed well~~ all the ML methods performed with minimum Root Mean Square Errors (RMSE) of about 0.1°C at a 1-day lead time and maximum values of about 0.8° ~~Cat~~ C at a 7-day lead time. In all regions, both the RForest and LSTM models consistently outperformed the CNN model across all lead times. LSTM has the highest predictive skill in 11 regions at all lead times. Importantly, the ML techniques ~~outperform~~ show results similar to the dynamical Copernicus Mediterranean Forecasting System (MedFS) for both SST and MHW forecasts, especially in the early forecast days. For MHW forecasting, ML methods ~~outperform~~ compare favorably with MedFS up to 3-day lead time in ~~most~~ 14 regions, while MedFS shows superior skill at 5-day lead time in 9 out of 16 regions. All methods ~~in all regions~~ predict the occurrence of MHWs with a confidence level greater than 50% in each region. Additionally, the study highlights the importance of incoming solar radiation as a significant predictor of SST variability along with SST itself.

1 Introduction

Accurate predictions of sea surface temperature (SST) and its extremes are important for many aspects of modern society. Anticipated changes include a rise in the occurrence and severity of prolonged Sea Surface Temperature extremes lasting a minimum of five days, commonly known as marine heatwaves (MHWs, Hobday et al. (2016)). These shifts have the potential to exert greater pressure on marine organisms and ecosystems, testing the boundaries of their adaptability and resilience ~~. This heightened stress could potentially result in irreversible harm to these ecosystems, as highlighted by studies conducted by Frölicher et al. (2018); Garrabou et al. (2022) and Garrabou et al. (2022).~~ (Frölicher and Laufkötter, 2018; Garrabou et al., 2022) . MHWs can affect marine biodiversity ~~(Garrabou et al., 2022; Cramer et al., 2018; Marbà et al., 2015; Rivetti et al., 2014; Juza et al., 2022~~ (Garrabou et al., 2022; Cramer et al., 2018; Marbà et al., 2015; Rivetti et al., 2014; Juza et al., 2022) and the fishing and aquaculture industries ~~(Cavole et al., 2016; Chandrapavan et al., 2019), by inducing mass-mortality, disease, large-scale coral bleaching and reduced seagrass meadows (Holbrook et al., 2020). Specifically, the Mediterranean Sea is a well-studied hot spot for MHW events (Garrabou et al., 2009; Giorgi, 2006; Cramer et al., 2018; Pastor et al., 2020; Pastor and Khodayar, 2022; Garrabou et al., 2022; Cia~~ ~~.~~ (Cavole et al., 2016; Chandrapavan et al., 2019). Thus, SST prediction, and in turn MHW prediction, can support a range of adaptive and management activities for the Mediterranean marine ecosystems.

Forecasting the anomalous oceanic and atmospheric patterns that drive the SST variability in the build-up to these extreme events is still a challenge (Jacox et al., 2022; Holbrook et al., 2020). In the last decades, dynamical ocean forecasting systems have increased spatial resolution and improved data assimilation techniques, simulating the dynamics of the global ocean down to a few kilometres, with the aim of kilometric resolution in their future generations (Leroux et al., 2022). Dynamical ocean predictions have reached a remarkable degree of reliability, although the required computational resources are enormous (Alvarez Fanjul et al., 2022). In recent years, increasing interest has been given to machine learning (ML) techniques, even though, in contrast to the dynamical model, ML techniques do not know when they are violating the laws of physics (Buizza et al., 2022). As a "learning from data" approach, machine learning has the advantages of computational efficiency, accuracy, transferability, flexibility and ease-of-use in ocean forecasting studies ~~(Boukabara et al., 2019; Li et al., 2020; Wei and Guan, 2022;~~ (Boukabara et al., 2019; Li et al., 2020; Wei and Guan, 2022; Taylor and Feng, 2022). Moreover, it is also less prone to model bias errors (Jacox et al., 2020) ~~. Machine~~ and, beyond computational efficiency, ML techniques excel in approximating nonlinear functions (Hornik, 1991).

Therefore, machine learning provides new opportunities for SST prediction (Boukabara et al., 2019). ~~For instance, the numerical approach is better suited for predictions over a wider area, while the data-driven techniques are more applicable for location-specific studies (Sarkar et al., 2020).~~ In contexts such as fishing, sporting events, coral bleaching and aquaculture, the ~~relevance of site-specific information is crucial. In these cases the~~ SST prediction can be treated as a time-series regression problem, where SST prediction is either restricted to a few locations or applied to SST values averaged over a region (Haghbin et al., 2021).

Machine learning techniques include both shallow methods, such as linear regression and Random Forest (RForest) models, and deep learning models, such as Artificial Neural Network (ANN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). As widely reviewed by Haghbin et al. (2021), linear regression and statistical methods ~~, such as the Autoregressive Integrated Moving Average (ARIMA) models,~~ have historically been extensively applied to SST estimation (Anding and Kauth, 1970; Corchado, 1995; McMillin, 1975). Pioneering works on SST prediction using deep learning methods are, instead, more recent. ~~Corchado and Fyfe (1999) applied a Negative Feedback ANN in different regions to predict time-series of SST, while Tangang et al. (1997) predicted sea surface temperatures over the Equatorial Pacific region~~

~~using a type of ANN, the multilayer perceptron (MLP). A similar investigation was carried out by Wu et al. (2006), who forecasted SST anomalies across the Tropical Pacific. ANN, such as MLP, were widely adopted at the begging of this century (Tang et al., 2000; Hsieh, 2000; Garcia-Gorriz and Garcia-Sanchez, 2007; Aguilar-Martinez and Hsieh, 2009).~~ In the last few years, as reported in a recent review paper (Haghbin et al., 2021), the deep learning-based models such as the RNN, in particular the Long ~~Short-Term~~ Short Term Memory (LSTM~~), and CNN~~ , e.g. Xiao et al. (2019); Liu et al. (2018); Xie et al. (2019) , and CNN (e.g. Han et al. (2019)) have attracted progressively more attention in the research community~~. For instance, Xiao et al. (2019), Liu et al. (2018) and Xie et al. (2019), who compared machine learning methods to predict SST in several different ocean areas, found LSTM to provide the most~~ , providing accurate estimates among the models considered. ~~An additional study involving LSTM networks was carried out by Wolff et al. (2020), who employed LSTM , Generalized Additive Model, RForest, MLP for SST modeling in the North Atlantic. Results demonstrate comparable performance of all the ML techniques to physics-based model simulations. Another study, by Han et al. (2019), found the CNN suitable for estimating SST in the Pacific ocean.~~ It is important to note that within the methods explored in this paper, LSTM stands out as the sole technique explicitly crafted for handling time series data. In contrast, CNN algorithms are commonly tailored for image processing tasks and are not inherently designed for time-series forecasting.

To the best of our knowledge, just a few studies employed the RForest model to predict sea surface temperature (Wolff et al., 2020), but, it has recently been successfully employed by Giamalaki et al. (2022) to directly predict marine extremes. Other attempts to predict extremes using ML techniques have been proposed by Chattopadhyay et al. (2020) and Jacques-Dumas et al. (2022) for land-based heatwaves over North America and France, respectively. Nevertheless, the advantage to have the SST prediction, instead of the MHWs events, is that end-users or management operators could establish thresholds based on their needs (Jacox et al., 2022).

In short, given the impacts of MHWs on ocean ecosystems and the resulting economic losses of marine industries, there is an increasing need for MHW forecasts to help ocean users to be prepared for these events. Short-duration extreme marine heatwaves (MHWs), lasting only a week, can exert sudden stresses on temperature-sensitive aquatic life stages, potentially leading to immediate adverse effects such as mass mortality, disease, large-scale coral bleaching and reduced seagrass meadows (Holbrook et al., 2020). For proactive marine management, operators in the coastal and marine sectors (e.g. fisheries and aquaculture, ~~shipping~~ and coastal water management) can use MHW event predictions for better planning of their activities. Here, we provide a proof-of-concept study on the advantage of data-driven ML methods to forecast the evolution of the SST state and its extremes in Mediterranean Sea regions up to 7 days ahead. The ML methods are compared against the Copernicus Mediterranean Forecasting System (MedFS, i.e. dynamical ocean model, (Clementi et al., 2021). MedFS is a numerical ocean prediction system, implemented and developed by the Euro-Mediterranean Center on Climate Change (CMCC), that produces analyses and short term forecasts for the entire Mediterranean Sea and adjacent areas in the Atlantic Ocean (Clementi et al., 2021). The Mediterranean Sea is a well-studied hot spot for MHW events (Garrabou et al., 2009; Giorgi, 2006; Cran . As detailed in the study by Bonino et al. (2023), the Adriatic Sea, the Gulf of Lion, and the Alboran Sea encountered the highest occurrence, shortest duration, and most severe MHWs. The Mediterranean Sea serves as our study area due to its relevance for marine management activities, indeed more than 95% of the global production of seabream and seabass comes from aquaculture, of which 97% is produced by Mediterranean countries (Carvalho and Guillen, 2021). To the best of our knowledge, this is one of the first attempts to predict SST and, in turn, MHWs in the Mediterranean Sea one week ahead using machine learning techniques.

This paper is organized as follows: in Sect. 2, we describe the methodological framework to build the machine learning techniques to predict SST and, in turn, MHWs. Sect. 3 reports the results and the comparison with a dynamical model. Our conclusions and outlook of the work are summarized in Sect. 4.

2 Methodological Framework

In this comprehensive study, the focus is on predicting SST time series and MHWs in 16 regions of the Mediterranean Sea (Figure 1) using ML techniques. In the following section we present the workflow of this study, which is summarised in Figure 2.

2.1 Data collection and preprocessing

The machine learning techniques are trained, tested, and validated using the European Space Agency (ESA) Climate Change Initiative SST dataset v2.1, referred to as the ESA CCI SST dataset in the following text (Merchant et al., 2019). This dataset, accessible through the CEDA catalogue [1], offers global daily satellite-derived Sea Surface Temperature (SST) data. ~~It covers~~ This dataset consists of daily maps of average SST at 20cm nominal depth with $0.05^{\circ}\times0.05^{\circ}$ of horizontal resolution, covering the period from September 1981 to December 2016. Additionally, to expand the temporal coverage, we incorporate daily Sea Surface Temperature data from 2017 to 2021, available via the Copernicus Climate Data Store (CDS)[2]. The extended data from 2017 to 2021 is generated at level L4 by the Copernicus Climate Change Service (C3S), building upon the foundation of the ESA CCI SST dataset. These datasets are derived using software and algorithms developed within the framework of the ESA CCI SST project. For a comprehensive insight into the updates in processing for the ESA CCI SST dataset v2.1, detailed information is provided in the work by Merchant et al. (2019).

The relevant atmospheric variables (AtmV) are taken from European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 dataset (Hersbach et al., 2020). Specifically, we select Sea Level Pressure (SLP), Geopotential Height at 500hPa (GEO), Wind Speed (WS), Sensible Heat flux (SENS), Latent Heat flux (LAT) and incoming ~~solar radiation~~ Solar Radiation (INC) as the sum of the short and long waves radiations downwards (i.e. into the ocean). SST and AtmV are averaged over the Mediterranean regions (Figure 1) to obtain time-series of SST and AtmV from 1981 to 2021. Moreover, we also consider the months of the year (MM) as an input variable, in order to describe the seasonality. Before building SST prediction tools based on machine learning, we analyze the mutual information between SST, SST itself and AtmV at different time lags (i.e. days) to have insights on the most relevant variables that contribute to the SST prediction. Mutual information between two random variables assesses their interrelationship. It signifies the amount of information that can be gained from one variable

[1] https://catalogue.ceda.ac.uk/uuid/62c0f97b1eac4e0197a674870afe1ee6~~last accessed: 14 March 2023~~

[2] https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-seasurfacetemperature?tab=overview

by observing another. A higher mutual information value corresponds to a greater reduction in uncertainty. Conversely, when mutual information is zero, the two variables are considered independent and unrelated. The mutual information of variable X and variable Y is defined as:

$$MI(x,y) = \int_x \int_y p_{XY}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \tag{1}$$

where $p_{XY}(x,y)$ is the joint probability density of X and Y, $p_X(x)$ and $p_Y(y)$ are the marginal probability density of X and Y, respectively. ~~Table 1 shows the mutual information between SST and the selected variables not lagged in time (LAG0) and with a 7 days lag (LAG7). In all the regions and for all the variables (except some rare cases for LAT) the mutual information decreases when increasing the time lag. In all the regions and for both the time lags, the INC, followed by the GEO, are the variables that seem to show the strongest association with SST, with values that range from 0.55 to 0.87 and from 0.38 to 0.59, respectively. INC influences directly SST during daily time, while GEO impacts ocean variability as a proxy for the large-scale atmospheric circulation. Increased GEO values correspond to higher air pressure above the ocean, resulting in warmer SST conditions. Progressively, we find MM, SLP and SENS with mean values of about 0.20, 0.15 and 0.10, respectively. WIND and LAT show low values, usually less than 0.1. However, it is worth noting that these values are substantially lower than those obtained considering the dependence between SST and itself at different lag (see Figure S1, reporting the mutual information heatmaps for all the time lags in each region). Despite the aforementioned values, reduced air-sea heat fluxes, in particular the latent heat, and reduced wind speed have been associated with MHWs in the mid-latitudes (Vogt et al., 2022). Therefore, we decide to retain all the selected AtmV as potential predictors of SST variability over the Mediterranean Sea.~~

2.2 Machine Learning Techniques:

Here, we briefly introduce the three ML techniques used in this study: a Long Short-Term Memory network (LSTM), a ~~one-dimensional~~ one dimensional Convolutional Neural Network (CNN) and a Random Forest model (RForest). While ~~a summary of the architectures is provided, more detailed descriptions can be obtained from~~ we offer a concise overview of their architectures here, for comprehensive details, readers can refer to Breiman (2001) for RForest, and Haghbin et al. (2021) for LSTM and CNN. All ~~the ML techniques applied in this study are multistep and multivariatemodels. They predict seven time-steps of SST into the future using multiple variablesas inputs~~these ML models are multifaceted and multivariate, projecting SST for seven subsequent time steps by utilizing various input variables. The input sequences ~~are the SST itself and the selected AtmVvariables, defined in the previous subsection, over~~ encompass SST data and the specified Atmospheric Variables (AtmV) detailed in the preceding section, spanning the previous 7 days. To ensure comparability across diverse atmospheric and oceanic variables, we normalize this data using a min-max scaler. These ML frameworks receive concatenated extensive vectors structured in the format [time-steps, lags, variables]. Here, 'lags' denotes the previous 7 days, while 'variables' encompass SST and the predictors. We develop LSTM and CNN architectures using Keras high-level API of the TensorFlow platform built in python [3], while Random Forest models using RandomForestRegressor function of sklean package of Python [4].

- **Long Short-Term Memory:** LSTM networks are types of recurrent neural networks capable of learning order dependence in sequence prediction problems, and they have been widely applied in temperature forecasting problems (Haghbin et al., 2021; Tran et al., 2021; Guo et al., 2022). We define the LSTM with 60 neurons with a hyperbolic tangent activation function in the first and unique hidden layer and 7 neurons in the dense layer (i.e. output layer) for predicting SST. Mean square error is used as loss function. The network was trained for 200 epochs using the Adam optimizer with a learning rate of 0.0001 and batch size of 150.

- **Convolutional Neural Network:** CNNs have gained significant popularity in domains like Image Processing and Computer Vision. In recent times, there has been a noticeable surge in interest within the research community to use CNNs for solving time-series forecasting problems. We define the one-dimensional CNN with 64 channels with a kernel size of 2 and with Rectified Linear Activation function in the first and unique hidden layer. It is followed by a maxpool layer which divides data size by 2. Finally a flatten operation and 7 neurons in the output layer followed. Mean square error is used as loss function. The network was trained for 200 epochs using the Adam optimizer with a learning rate of 0.0001 and batch size of 150.

- **Random Forest Model:** RForest is an ensemble learning method for classification and regression tasks that operates by constructing a multitude of randomly-perturbed decision trees starting form the same train set. For regression tasks, the mean or average prediction of the individual trees is returned. We design the RForest model with 100 decision trees and 42 predictors randomly selected to perform each split to construct decision trees. The function to measure the quality of a split during training is the mean squared error.

Notably, in our methodology, we opted not to incorporate dropout layers within the network architecture. Despite the absence of dropout regularization, our training results did not exhibit overfitting tendencies, suggesting a congruence between the model's capacity and the dataset characteristics.

2.3 Experiments

36 years of the daily data of SST and atmospheric variables are used to train and validate the techniques, while the remaining 4 years are used to test it. Based on the years used to train and test the methods we distinguish two kinds of experiments:

- Reference Experiments (*REXPs*): ML techniques are trained using 1981-2016 as the training period and 2017-2021 as the testing period. We are interested in predicting the future SST, so the methods have to learn the correct time evolution of the SST and, moreover, we also want to compare the skill with the ~~Copernicus Mediterranean Forecasting System (MedFS , i.e. dynamical ocean model, (Clementi et al., 2021)). MedFS is a numerical ocean prediction system,~~

[3] https://www.tensorflow.org/api_docs/python/tf/keras

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

~~implemented and developed by the Euro-Mediterranean Center on Climate Change (CMCC), that produces analyses and short term forecasts for the entire Mediterranean Sea and adjacent areas in the Atlantic Ocean (Clementi et al., 2021). MedFS predictions are operationally delivered at~~ MedFS dynamical model. MedFS predictions, part of the Copernicus Marine Service since 2017, offer SST forecasts for lead times up to 9 days averaged across specified regions. Since the accuracy of the ML forecast decreases almost linearly with time, we decided to limit the comparison with the MedFS forecasted SST averages for lead times of 1~~/24 degree horizontal resolution since 2017. In this contex we use MedFS forecast SST data with 1~~, 3, and 5 days~~lead time~~.

- Uncertainties Experiments (*UEXPs*): the sampling uncertainty of each ML technique ~~has been~~ are estimated by progressively sliding the training and the testing period window of 4 years from 1981 to 2017. This means that, for each region, we perform 14 uncertainty experiments. For example the first experiment uses 1985-2021 as training period and 1981-1984 testing period, while the second one merges together 1981-1984 and 1989-2021 data to create the training dataset and it uses 1984-1988 data as testing dataset and so on.

~~The computational time required is about 30 minutes to train the method and 1 minute to test it.~~ In addition, we investigate the role of each driver in affecting prediction skills. Thus, for each ML technique and for each experiment (*REXPs* and *UEXPs*), the prediction accuracy on the test set was evaluated after randomly permuting (shuffling) the value of each driver, one at time (i.e. one "experiment" for each shuffled driver). This random permutation is aimed at removing any information about SST conveyed by the drivers (that is, annihilating the mutual information between SST and the driver), thus voiding their contribution in predicting SST (Breiman, 2001). In case of informative drivers, the shuffling is expected to severely affect prediction accuracy, while it should have a negligible impact when considering uninformative drivers. In the following we refer to these sensitivity experiments as *SEXPs*.

The computational time required for the ML experiments is about 2 minutes to train the method and 30 seconds to test it, while the MedFS requires 10 minutes to simulate 1 day (i.e. 70 minutes for 7 days of forecast). We run all the experiments on the CMCC supercomputer ZEUS, which comprises 348 Lenovo SD530 biprocessor nodes (totaling 12,528 cores) interconnected via an Infiniband EDR network and boasts a total theoretical computing power of 1,202 TFlops.

2.4 Evaluation metrics

For evaluating the ML-based prediction skills, we use ~~of~~ a metric that is commonly applied in the SST forecast domain. In particular, we calculate the root mean square error (RMSE, Eq. 2) of the predicted SST in the test datasets against the ESA CCI SST dataset (i.e. observed SST). The RMSE measures the mean squared distance between the daily predicted (F_i) and the daily actual (T_i) SST in the N samples of the test dataset. The RMSE is negatively oriented with a perfect value of 0.

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{n}(T_i - F_i)^2} \tag{2}$$

Moreover, we also assess the ML techniques~~’~~’ accuracy in predicting MHW occurrence by detecting MHWs in the test time-series and in the predicted time-series. Notably, the test time-series represents the regional mean SST, resulting in a single time-series per region. MHWs are defined ~~as in Hobday et al. 2016 (Hobday et al., 2016)~~in each regional time-series as in Hobday et al. (2016): SST higher, for 5-days or longer, than the 90th-percentile threshold of seasonally varying climatology calculated over more than 30-year period without removing long-term trend. If an event occurs less than 2 days before the end of another, it is considered as part of the ongoing MHW. An 11-day sliding window centered on each day is used for the climatology calculation. The climatology and the threshold are calculated over the reference ~~testing~~ training period (i.e. ~~1981-2016~~1982-2016) and applied to the ~~test~~ testing period (i.e. 2017-2021). Even though MHWs are defined considering at least 5 days of consecutive anomalies, the differences between the predicted and observed datasets are calculated day by day. The detection performance is assessed by computing, in the test set, the rates of False Positive (FPR, Type I error: incorrect detection of MHW) and False Negatives (FNR, Type II error: non-detection when MHW occurs) and the F1 score. FPR and FNR are defined as:

$$FPR = \frac{FP}{FP+TN} \quad \text{and} \quad FNR = \frac{FN}{FN+TP} \tag{3}$$

where FP is the number of False Positives, TN is the number of True Negatives, FN is number of False Negatives and TP is the number of True Positives (see Table 3). Note that the True Positive Rate (TPR) and the True Negative Rate (TNR) are $TPR = 1 - FNR$ and $TNR = 1 - FPR$, respectively. The F1 score is a single overall measure of prediction accuracy (Eq.4), and takes into account the imbalance of the dataset: around 1/7th of the test samples are MHWs events. It is calculated from the Precision and Recall scores.

$$F_1 = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{where} \quad \text{Precision} = \frac{TP}{(TP+FP)} \quad \text{and} \quad \text{Recall} = \frac{TP}{(TP+FN)} \tag{4}$$

We also assess the mean forecasted intensity ($Imean_{ML}$) within the *REXPs* for each ML technique and compare it to the observed intensity mean from the ESA CCI dataset ($Imean_{ESA}$), computing the Intensity Error (IE), formulated as $IE = Imean_{ESA} - Imean_{ML}$. Furthermore, we compare the performance of the SST and MHW prediction of the *REXPs* against MedFS SST forecast data.

3 Results

In the following, we evaluate the predictive skill of the ML techniques for SST predictions and, in turn, MHW predictions, up to a lead time of 7 days. For the presented results, the statistical index used to evaluate the performances is the RMSE for SST and the F1 scores for MHWs (see "Evaluation metrics" section). Moreover, we also assess the sensitivity of the ML techniques with respect to the input variables in the testing dataset.

3.1 Mutual information analysis

Table 1 shows the mutual information between SST and the selected variables not lagged in time (LAG0) and with a 7 days lag (LAG7). In all the regions and for all the variables (except some rare cases for LAT) the mutual information decreases when increasing the time lag. In all the regions and for both the time lags, the INC, followed by the GEO, are the variables that seem to show the strongest association with SST, with values that range from 0.55 to 0.87 and from 0.38 to 0.59, respectively. INC thermally influences the temperature along the water column, while GEO impacts ocean variability as a proxy for the large-scale atmospheric circulation. Increased GEO values correspond to higher air pressure above the ocean, resulting in warmer SST conditions. Progressively, we find MM, SLP and SENS with mean values of about 0.20, 0.15 and 0.10, respectively. WIND and LAT show low values, usually less than 0.1. However, it is worth noting that these values are substantially lower than those obtained considering the dependence between SST and itself at different lag (see Figure S1, reporting the mutual information heatmaps for all the time lags in each region). Despite the aforementioned values, reduced air-sea heat fluxes, in particular the latent heat, and reduced wind speed have been associated with MHWs in the mid-latitudes (Vogt et al., 2022). Therefore, we decide to retain all the selected AtmV as potential predictors of SST variability over the Mediterranean Sea.

3.2 SST prediction

To have an overall insight of the methods~~’~~’ performances over the Mediterranean Sea, we first examine the ML methods~~’~~’ performance of the daily SST prediction in all the regions for the *REXPs* (Figure 3). Overall, these methods show similar ranges of RMSE, they display minimum values of about 0.1°C at lead time of 1 day (L1) and maximum values of about 0.8°C at lead time of 7 days (L7). The RMSEs grow with the increasing forecast lead time, and the evolution of the RMSE in each region is consistent across ML techniques. For instance, in all the techniques, region 15 shows the lowest, or almost the lowest (e.g in CNN) RMSE, while region 11 shows the highest errors. To visualize the errors spatially, Figure 4 shows the mean RMSE for LSTM for each region (CNN and RF results are shown in Figure S2). The areas with larger errors are: (i) the Alboran Sea, strongly characterized by a complex dynamics influenced by the incoming cold Atlantic water through the Gibraltar Strait modulating the water transport; (ii) the North-West part of the basin, which is an area of dense water formation and intense dynamics due to the Gulf of Lion gyre (Madec et al., 1991; Pinardi et al., 2006) and by a boundary intense current called the Liguro-Provenal-Catalan Current (Pinardi et al., 2006); (iii) the Adriatic Sea, especially in its northern shelf, characterized by a complex topography, intense air-sea exchanges, large riverine inputs that contribute to enrich the dynamics of the area. Looking more in detail into the methods~~’~~’ performance comparison (Table 4), LSTM, followed by RForest, outperforms CNN ~~both~~ at L1 and at L7. In particular at L1, LSTM has the highest predictive skill in ~~8~~ 13 out of 16 regions, while RForest in ~~5~~ 2 out of 16 regions. In the remaining ~~4 regions~~region, they score equally~~for region 9 at both lead times, while for region 2,~~ . For L7, LSTM has the highest predictive skill in 11 out of 16 regions, while RForest in 1 out of 16 regions. In the remaining 4 ~~and 14 RForest (LSTM) shows lower RMSE values at L1 (L7).~~

regions, they score equally.

We select regions 11, 15, 4 which display the highest, the lowest, and an intermediate RMSEs, respectively. ~~Moreover, different techniques have the best performance in these areas: LSTM for region 11, RForest for region 15, and RForest (LSTM) at L1 (L7) for region 4. They also represent different dynamical areas of the Mediterranean basin.~~ We refer to region 4 as “Western Mediterranean, WM”, to region 11 as “Central Mediterranean, CM” and to region 15 as “Eastern Mediterranean, EM”. Figure ~~5a~~ 4a shows the RMSEs of the predicted SST by *REXPs* (solid line) and by *UEXPs* (bars) against the observed SST time-series.

Moreover, the RMSEs of the SST predicted by MedFS (i.e. dynamical model) for the reference period (i.e. ~~2017-2021~~2017-2021) are also reported. It is worth noting that the RMSE of the dynamical model does not show significant increases with the forecast lead time, unlike ML techniques. ~~In contrast to ML methods, the dynamical model's prediction of SST is influenced by atmospheric forecasts throughout the forecast period, which likely prevents the RMSE from increasing with the lead time.~~ Results for all the other regions are shown in Figure ~~S2~~S3. Referring to Figure 5a and Figure ~~S2~~S3, we can appreciate that most of the ML techniques errors compare favorably with respect to the MedFS errors during the first days of forecast. In particular, the CM RMSEs range between a minimum error of about ~~0.21~~0.19°C at L1 and maximum error of about ~~0.78~~0.74°C at L7. All the ML methods show lower RMSE than the MedFS forecast system for the first 3 days of forecast and they are comparable at lead time of 5 days. EM and WM show lower variability of the error with respect to the CM, ranging in the intervals ~~0.13~~0.11°C - ~~0.45~~0.35°C and ~~0.16~~0.14°C - ~~0.51~~0.45°C , respectively. Over those two regions, it could be observed that all ML methods~~'~~' skills are in line with the one of MedFS, they have similar RMSEs, being CNN the one showing higher error~~and larger uncertaintiy (indicated by the length of the bars) with respect to LSTM and RForest models~~. In WM and ~~EM~~ CM - and in almost all the other regions (see Figure ~~S2~~S3) - the uncertainties tend towards higher RMSEs with respect to the *REXPs* errors (i.e. errors represented by the solid line in Figure 5a). It is likely connected to the fact that CNN algorithms are typically designed for image processing rather than time-series forecasting.

An additional analysis is presented to show how the different ML methods perform in predicting SST and MHWs occurrence at different forecast lead time. Figure 6 shows 1 year (2020) daily SST time-series of the predicted and observed SST at L5 (Figure ~~S3 and~~ S4 and S5 for L1 and L3 respectively) as well as the SST climatology, averaged in the 3 regions of interest. The figure shows a very close match between the forecasts and the observations, being the SST variability clearly well represented and forecasted by all the models (i.e. ML techniques and MedFS model). This is confirmed by the fact that the difference in annual means and standard deviations are minimal, generally within decimal values.

3.3 MHWs prediction

Going a step further in the prediction skill assessment, we also evaluate the ability of the different ML techniques in predicting MHWs occurrence (Table 5 and Figure 5b). To this end, we define MHWs using the method of ~~(Hobday et al., 2016)~~ Hobday et al. (2016) in the observed time-series, in the predicted time-series by ~~*REXPs*~~ REXPs and by MedFS (see "Evaluation metrics" section for more details). For the selected regions, Table ~~5~~ 4 reports the False Positive Rates (FPR), the False Negative Rates (FNR) and the F1 scores at L1, L3 and L5~~, while Figure 5b shows the variation of the F1 score for all the methods with increasing forecast lead time~~. Results for all the other regions are reported in Figure S5 and Table S1. Overall for all the forecast lead time and for all the ML techniques (except in rare cases, see Table S1), the FNR is higher than the FPR, meaning that the ML methods tend to underestimate SST peaks/extremes. MedFS model, instead, shows mixed behaviour: 7 out of 16 regions show higher FPR than FNR at all the lead times. In CM, the MedFS shows a high FPR of about 27% at L5 (Table 5) as it is also evident in Figure 6. During January 2020 and 2021 the MedFS predicted time-series (blue line in Figure 6) is usually greater than the 90th percentile threshold used to define MHWs (gray line in Figure 6), leading to high FP (highlighted as circles in Figure 6). The ML techniques show, instead, high FNR of about 50% (highlighted as crosses in

Figure 6). Nevertheless, in WM, MedFS is in line with ~~the ML methods, except for LSTM ,~~ LSTM and RForest showing 30% of FNR. Looking from another perspective, and reminding that $TPR = 1 - FNR$, these rates demonstrate that, except for some rare cases, all the methods in all the regions predict the occurrence of a MHW with a confidence greater than 50%.

To have a more accurate evaluation of the performance of the methods, instead of showing just the errors (i.e. FPR and FNR) we also evaluate the F1 score (Figure 5b and Table5 and S1). Note that the F1 score ranges between 0 and 1 ~~an~~ and it is positively oriented. As expected, the F1 scores decrease when increasing the forecast lead time, but, at all the lead times, they show values greater than 0.5, meaning satisfactory MHWs predictions for all the ML techniques in all the regions (except some rare cases, especially for CNN). Overall, comparing the performances of the ML techniques, the RForest and the LSTM outperform CNN in all the regions, at least at the first days of forecast. The F1 score results show that ~~RForest~~ LSTM outperforms in the ~~EM~~ CM and WM at all the lead times reaching a F1 score of about 0.9 at L1 (Figure 5b). ~~WM and CM show~~ EM shows different behaviour: ~~RForest~~ LSTM has the best predictive skill, outperformed by ~~LSTM~~ RF in the following lead times. Nevertheless, as for the SST RMSE, all the ML techniques provide nearly same results; indeed, the ML differences in terms of F1 score are usually around 0.15. Comparing the results with MedFS model (blue circles in Figure 6) we can appreciate that in the selected regions the ML techniques outperform the MedFS up to L5 for ~~WM and~~ CM, and up to L3 for WM and EM (Table 5). It is worth noting that in all the other regions (except region 8) ML methods outperform the MedFS up to L3, while at L5 in 9 out of 16 regions the MedFS has the best skill (Table S1).

We also evaluate the differences in MHWs intensity mean predicted by the ML models and MedFS against observation (i.e. intensity error, IE, see "Evaluation metrics" section) during the studied period (Figure 5c). For WM, CM, and EM it is interesting to note that MedFS shows positive IEs, meaning that the MedFS predicted intensity is overestimated, while the ML models show opposite behavior, IEs are negative. This characteristic is evident also for the other regions (Figure S6), except regions 1, 12, 13 where MedFS IEs are negative. It is also worth highlighting that, in all the regions, the RForest IEs, even if they are not the smallest ones, degrade slower than the LSTM IEs.

3.4 Sensitivity analysis

In this section, we discuss the results of the *SEXPs* (see "Experiments" section for detail) for the ~~selected 3~~ 3selected regions, noting that similar conclusions can be drawn for all the regions (not shown). The analysis thereby focuses on evaluating the methods~~'~~' performance in terms of SST RMSE (Figure 7). This means that the higher the increase in the RMSE after a driver is shuffled, the higher its predictive power. The labels of Figure 7 indicate, for each experiment, the driver that has been shuffled in *REXPs*. For an easier interpretation of the results we are showing only the SST, the INC and the LAT drivers, as they are the most relevant. ~~They also show the *REXP* errors, i.e. solid line in Figure 5.~~ For all the techniques, the RMSE increases notably with respect to the *REXP* when the SST is randomly modified, it grows up to 6°C, 7°~~Cand~~ C and 5.5°C for WM, CM and EM, respectively. Nevertheless, it is worth noting that the extent to which the RMSE increases after shuffling SST shows a tendency to decrease as the forecast lead time increases. This result suggests that the SST itself has the strongest predictive power in forecasting SST, slightly losing predictive skill increasing the lead times. The incoming solar radiation, to a lower extent, shows the opposite behaviour: after shuffling, the RMSE tends to increase more than the other drivers with the forecast lead times. The RMSE at L7 reached values of about 1°~~C, 1.5~~CC, 1.7°~~C, 0.75~~CC, 1°~~C~~ CC for WM, CM and EM, respectively. Surprisingly, in contrast with the mutual information analysis, ~~for CNN~~ we can notice that for WM and EM the latent heat plays a role~~in the first days of forecast, being outperformed by the incoming solar radiation at the following lead times. Evaluating this result, we should take into account that CNN showed the lowest performance both in terms of SST and MHWs (see Figure 5, Figure S2, Figure S5 and Table S1)~~. In particular, for WM the RMSE at L7 reached values of about 0.5 for all the ML techniques, double of the RMSE of *REXPs*. Overall, the aforementioned analysis suggests that the incoming solar radiation, as shown also by the mutual information analysis, has some predictive power in driving SST variability. It is important to highlight that incoming solar radiation shows a tendency to gain predictive power as forecast leads increase, whereas SST, to a much lesser extent, tends to lose it. This suggests that atmospheric variables could be useful in forecasting SST at longer time scales. ~~Nevertheless~~However, it is worth ~~mentioning~~ stressing that the ML methods look for statistical relations (e.g. linear or non-linear relations) between variables that do not ~~have necessary~~ necessarily have a physical meaning (e.g. a cause-effect relation).

4 Discussion~~and Conclusions~~

In this study, a group of ML algorithms - Random Forest (RForest), Long short-term memory (LSTM) and Convolutional Neural Networks (CNN) - are used to evaluate their ability in building a competitive prediction tool of ~~Sea Surface Temperature (SST) prediction and Marine Heatwave (MHW)~~ SST and MHWS occurrence 7 days ahead in the Mediterranean Sea. The methods use the European Space Agency (ESA) Climate Change Initiative (CCI) Sea Surface Temperature, Sea Level Pressure (SLP), Geopotential Height at 500hPa (GEO), Wind Speed (WS), Sensible Heat flux (SENS), Latent Heat flux (LAT) and incoming solar radiation (INC) from ECMWF ERA5 as input data.

~~The results demonstrate that ML techniques attain some predictive skill in forecasting the Sea Surface Temperature evolution and the MHW occurrences over the Mediterranean Sea a week in advance, especially RForest and LSTM demonstrated to be accurate. RForest and LSTM exhibited similar ranges of RMSE, with minimum values of about 0.1°C observed at a lead time of 1 day and maximum values of approximately 0.8°C at a lead time of 7 days~~ We compare the ML predictions against the MedFS prediction system, part of the Copernicus Marine Service since 2017, which offers SST forecasts for lead times up to 9 days averaged across specified regions. It is ~~worth mentioning that among the methods considered in this paper, LSTM is the only one specifically devised to deal with time series data.~~ important to underline that the data used in our work are designed primarily for climate studies and providing a gap-free dataset through interpolation, raises concerns about potential biases introduced into our forecasting model. The interpolation process, while ensuring a comprehensive dataset, might inadvertently smooth variations and obscure critical phenomena like coastal upwelling. Alternative approaches utilizing near real-time operational data, could offer more dynamically responsive and less biased datasets for improved forecasting accuracy. Moreover, our methodology involves averaging SST data to obtain a single representative SST for each zone and the same approach is used for atmospheric predictors. We acknowledge that averaging data in this manner might smooth out localized variations, thereby potentially overlooking non-linearity in the effects of variables such as wind speed. When data are averaged within an area, the issue tends to become more linear, allowing simpler methods like linear regression to yield good results (not shown). However, when working with higher resolution data, ML methods outperform classical methods demonstrating their potential and unearthing intricate nonlinear relationships.

~~Regarding MHW forecasting , ML methods in all regions achieved a confidence level greater than 50% when predicting the occurrence of an MHW. Nevertheless, as also shown by Giamalaki et al. (2022), ML methods are characterized by high rate of True Positives and high rate of False Negatives. High rates of False Negatives suggest that the ML techniques tend to underestimate SST peaks/extremes. In addition, one has to keep in mind the imbalance characterising this prediction problem, in which the number of days with MHW conditions is substantially smaller than the number of days without MHW.~~

A crucial aspect of this study is the comparison of the ML techniques' performance with that of the dynamical Copernicus Mediterranean Forecasting System (MedFS) for both SST and MHW forecasts. Impressively, ML methods demonstrated a favorable edge over MedFS, especially LSTM in the early forecast days. For MHW forecasting, ML methods outperformed MedFS in most regions up to 3 days of forecast lead time, while MedFS exhibited superior skill at 5 days of forecast lead time in 9 out of 16 regions. It is worth emphasizing that dynamical models, used to produce the 7-day forecast, are forced by atmospheric forecasts, enabling the ocean to be influenced by the overlying atmospheric conditions. On the contrary, ML methods lack information about the atmospheric conditions during the forecasted period and heavily depend on the autocorrelation structure of SST, which rapidly decreases as the time lag increases. These could be the causes of the faster increase in RMSE with the forecast lead times compared to MedFS. Results demonstrate comparable performance, at least at the earliest days of forecast, to physics-based model simulations (i.e. Copernicus Mediterranean Forecasting System) but with the advantage of low computational cost. The low computational cost of these off-the-shelf ML tools has many advantages. First, the suite of methods presented here can be trained on a laptop and applied to any geographic location. Secondly, once trained, the ML techniques do not require high user skills to be correctly run and analysed. Furthermore, they can be easily updated, once additional data ~~become~~ becomes available. In addition, the advantage to ~~have~~ having the SST prediction is that the end-users could establish thresholds based on their needs. Marine users and stakeholders operating with different purposes and in different regions may need specific thresholds to define the extreme conditions which may limit their activities. However, the gap to dynamical models at 7 days of forecast lead time and the high rates of False Negatives motivate future work to improve the performance of the underlying networks. For example, one may consider adding complexity or improving the model architectures. Passing from time-series forecasts to spatially-complete maps of predicted sea surface temperatures (i.e. from 1D prediction to 2D prediction), is also appealing. These methods, contingent ~~to~~ on higher computational time and resources, could be trained in each grid point of the target dataset, in order to obtain maps of SST prediction for each lead time at very high resolution.

Our findings also indicate that, in addition to SST itself (as also observed by Giamalaki et al. (2022)), incoming solar radiation appears to play a role in predicting SST. These ~~two~~ variables are inherently physically related, but it is important to note that ML techniques, unlike dynamical models, do not simulate the ocean's dynamics. Therefore, establishing a physical process-based relationship between incoming solar radiation and MHW occurrences is premature. To comprehend this underlying connection, driver-based studies, such as those conducted by ~~Holbrook et al. (2019), Schlegel et al. (2021), and~~

~~Rodrigues et al. (2019), are necessary.~~ Holbrook et al. (2019); Schlegel et al. (2021); Rodrigues et al. (2019), are necessary. Furthermore, the neural network algorithms applied lack inherent knowledge of physical laws, potentially leading to violations of fundamental physical constraints. This limitation likely restricts the proficiency of NNs in maintaining accuracy throughout the prediction interval ((Boukabara et al., 2019; Dueben and Bauer, 2018). It prompts a need for cautious interpretation and validation of results, especially in scenarios where physical constraints significantly influence outcomes. While this study primarily focuses on predictive skill, future iterations could explore methodologies to infuse ML frameworks with physical laws governing ocean dynamics. For instance, Zanetta et al. (2023) propose achieving physical consistency in deep learning-based postprocessing models for temperature and humidity by incorporating meteorological expertise through analytic equations. Incorporating these constraints could potentially enhance the reliability of NN predictions, mitigating the risk of straying from physical realities. One may consider, due to the strengths of NNs in shorter-term predictions (3/5-days interval) and the potential limitations discussed, to use the ML model alongside models like MedFS for longer prediction intervals. Utilizing MedFS for longer-term forecasts could leverage its established reliability over extended periods, while ML could excel in shorter-term predictions.

Our study predominantly focuses on the influence of atmospheric forcings on SST predictions, motivated by their significant role in the onset of MHWs (Schlegel et al., 2021; Darmaraki et al., 2023). While this approach captures pivotal aspects, it fails to account for a variety of oceanic processes that could contribute significantly to SST variations and MHWs events. The drivers of MHWs are currently not fully understood (Holbrook et al., 2019), and the relevant physical drivers and processes involved in MHW emergence span various timescales, ranging from days (e.g., anomalous heat fluxes) to weeks (e.g., blocking systems and atmospheric teleconnections), months (e.g., re-emergence of warm anomalies from the subsurface), and years (e.g., climate modes and oceanic teleconnections). ~~Furthermore, it is worth emphasizing that dynamical models, used to produce the 7-day forecast, are forced by atmospheric forecasts, enabling the ocean to be influenced by the overlying atmospheric conditions. On the contrary, ML methods lack information about the atmospheric conditions during the forecasted period, which likely explains the faster increase in RMSE with the forecast lead times compared to MedFS. To address this, explicit utilization of gained physical knowledge and run-time conditions could be employed to construct model architectures. This could be achieved during preprocessing through feature engineering or during training by formulating physical constraints (de Bézenac et al., 2019; Karniadakis et al., 2021).~~ By omitting these factors, our study might provide an incomplete understanding of the intricate interplay between oceanic and atmospheric dynamics, limiting the comprehensiveness of our SST forecasting model. Introducing additional ocean variables alongside atmospheric data could enrich the predictive capacity of ML techniques, improving predictions by integrating a more comprehensive set of influencing factors

~~In brief, data-driven~~ Data-driven methods used to forecast SST and, in particular, MHW occurrence on a weekly basis, are still in their research infancy. In general, weekly MHW predictions are currently missing from the literature, although weekly forecasts of ocean conditions are widely available (Giamalaki et al., 2022). The presented work helps to demonstrate and confirm the power of these easy-to-use tools which could be efficiently applied to predict the future state of the ocean one week ahead. ~~Numerous industries can benefit from these~~ We recognize the merit of longer-term forecasting for understanding and mitigating lasting impacts, but short-term forecasts ~~, including fisheries, aquaculture farms and coastal water management~~ supply useful information in various sectors, specifically in industries reliant on immediate responses to temperature fluctuations (?). Short-term SST and MHW forecasts are particularly relevant in sectors such as aquaculture, a very important activity in the Mediterranean sea, where the acute impacts of short-duration extreme events can significantly impact operations (DeMott et al., 2021~~). Week-ahead alerts can assist fishery managers in avoiding potential closures and allow them to prepare the industry for necessary modifications in gear and labor. Aquaculture managers can take protective measures against excessive warming or~~ , Frölicher and Laufkötter (2018)). For instance, an extreme MHW lasting a week might impose sudden stresses on temperature-sensitive aquatic species (e.g. the early life stages of fish, Buttay et al. (2023)), potentially causing immediate adverse effects, such as mass mortality (Lee et al., 2018; Buttay et al., 2023; **?**) compared to a more prolonged, moderate MHW. Timely predictions of these short-term fluctuations enable proactive management strategies, aiding in disease prevention, optimizing feeding schedules, potential coral bleaching events ~~at aquaculture sites. Valuable information can be provided to make prudent decisions in advance of potentially severe economic catastrophes. Therefore, data-driven approaches to forecast~~ prevention, and ensuring the well-being of farmed species, hence safeguarding economic interests and sustaining food security (**??**).

5 Conclusions and future work

Our study is designed as a proof of concept to showcase the potential applications of machine learning methods in short-term (up to 7 days) SST and MHWs ~~could be seen as an alternative approach that would provide rapid early warning to resource managers and ocean stakeholders who could take action to reduce potential impacts on the ecosystem~~ forecasting. ML methods, especially LSTM, showed an early edge over dynamical models, marking a preliminary step toward advanced systems capturing non-linear connections. Future efforts should refine ML models by addressing data averaging limits and exploring higher resolution data. Integrating physical laws into ML frameworks could enhance reliability, while exploring additional oceanic variables would aid in understanding dynamics, especially MHW drivers. These forecasting tools offer proactive strategies for aquaculture and ecosystem management, providing rapid warnings for stakeholders to mitigate ecosystem impacts.

Code and Data Availability

The ~~codes~~ code generated during the current study ~~are available from the corresponding author on reasonable request.~~

~~Data Availability~~

to train and test the ML model is stored in Zenodo repository (**?**, https://doi.org/10.5281/zenodo.8335345).). All data used in this study is open access. The SST dataset used in this study is the European Space Agency (ESA) Climate Change Initiative SST dataset v2.1 (Merchant et al., 2019) and it is freely available at the CEDA catalogue here: https://catalogue.ceda.ac.uk/uuid/62c0f97b1eac4e0197a674870afe1ee6) from September 1981 to December 2016 and in the Copernicus CDS here: https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-seasurfacetemperature?tab=overview from January 2017 to December 2021. The relevant atmospheric variables are taken from European Centre for Medium-Range Weather Forecasts (ECMWF)

ERA5 dataset (Hersbach et al., 2020) at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview.

Author contributions statement

G.B, G.G., S.M. conceived the study. G.B, S.M and G.G. discussed and defined the methodological framework. G.B. performed the experiment and wrote the manuscript. G.B., S.M., G.G., interpreted the results. G.B., S.M., G.G.,R.M.,E.C. contributed to the interpretation of the results and to the paper writing. All authors reviewed the manuscript.

Additional information

Competing Interests: The authors declare no competing interests.

Acknowledgements. This research has been funded by the European Space Agency (ESA) as part of the FEVERSEA Climate Change Initiative (CCI) fellowship (ESA ESRIN/Contract No. 4000133282/20/I/NB). We acknowledge the CMCC Foundation for providing computational resources.

References

Aguilar-Martinez, S. and Hsieh, W. W.: Forecasts of tropical Pacific sea surface temperatures by neural networks and support vector regression, International Journal of Oceanography, 2009, 2009.

Alvarez Fanjul, E., Ciliberti, S. A., and Bahurel, P.: Implementing Operational Ocean Monitoring and Forecasting Systems., 2022.

Anding, D. and Kauth, R.: Estimation of sea surface temperature from space, Remote Sensing of Environment, 1, 217–220, 1970.

Bonino, G., Masina, S., Galimberti, G., and Moretti, M.: Southern Europe and western Asian marine heatwaves (SEWA-MHWs): a dataset based on macroevents, Earth System Science Data, 15, 1269–1285, 2023.

Boukabara, S.-A., Krasnopolsky, V., Stewart, J. Q., Maddy, E. S., Shahroudi, N., and Hoffman, R. N.: Leveraging modern artificial intelligence for remote sensing and NWP: Benefits and challenges, Bulletin of the American Meteorological Society, 100, ES473–ES491, 2019.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Buizza, C., Casas, C. Q., Nadler, P., Mack, J., Marrone, S., Titus, Z., Le Cornec, C., Heylen, E., Dur, T., Ruiz, L. B., et al.: Data learning: Integrating data assimilation and machine learning, Journal of Computational Science, 58, 101 525, 2022.

Buttay, L., Ohlberger, J., and Langangen, Ø.: Management strategies can buffer the effect of mass mortality in early life stages of fish, Journal of Applied Ecology, 60, 2056–2065, 2023.

Carvalho, N. and Guillen, J.: Aquaculture in the Mediterranean, IEMed Mediterr. Yearb, 2021.

Cavole, L. M., Demko, A. M., Diner, R. E., Giddings, A., Koester, I., Pagniello, C. M., Paulsen, M.-L., Ramirez-Valdez, A., Schwenck, S. M., Yen, N. K., et al.: Biological impacts of the 2013–2015 warm-water anomaly in the Northeast Pacific: winners, losers, and the future, Oceanography, 29, 273–285, 2016.

Chandrapavan, A., Caputi, N., and Kangas, M. I.: The decline and recovery of a crab population from an extreme marine heatwave and a changing climate, Frontiers in Marine Science, 6, 510, 2019.

Chattopadhyay, A., Nabizadeh, E., and Hassanzadeh, P.: Analog forecasting of extreme-causing weather patterns using deep learning, Journal of Advances in Modeling Earth Systems, 12, e2019MS001 958, 2020.

Ciappa, A. C.: Effects of Marine Heatwaves (MHW) and Cold Spells (MCS) on the surface warming of the Mediterranean Sea from 1989 to 2018, Progress in Oceanography, p. 102828, 2022.

Clementi, E., Aydogdu, A., Goglio, A., Pistoia, J., Escudier, R., Drudi, M., Grandi, A., Mariani, A., Lyubartsev, V., Lecci, R., et al.: Mediterranean Sea Physical Analysis and Forecast (CMEMS MED-Currents, EAS6 system)(Version 1)[Data set], Copernicus Monitoring Environment Marine Service (CMEMS), 10, https://doi.org/https://doi.org/10.25423/CMCC/MEDSEA_ANALYSISFORECAST_PHY_006_013_EAS6, 2021.

Corchado, J.: Hybrid cbr system for real-time temperature forecasting in the ocean, in: IEEE colloquium on knowledge discovery. London, UK, 1995.

Corchado, J. M. and Fyfe, C.: Unsupervised neural method for temperature forecasting, Artificial Intelligence in Engineering, 13, 351–357, 1999.

Cramer, W., Guiot, J., Fader, M., Garrabou, J., Gattuso, J.-P., Iglesias, A., Lange, M. A., Lionello, P., Llasat, M. C., Paz, S., et al.: Climate change and interconnected risks to sustainable development in the Mediterranean, Nature Climate Change, 8, 972–980, 2018.

Darmaraki, S., Waldman, R., Sevault, F., and Somot, S.: Dominant drivers of Past Mediterranean Marine Heatwaves, Tech. rep., Copernicus Meetings, 2023.

Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geoscientific Model Development, 11, 3999–4009, 2018.

Frölicher, T. L. and Laufkötter, C.: Emerging risks from marine heat waves, Nature Communications, 9, 1–4, 2018.

Frölicher, T. L., Fischer, E. M., and Gruber, N.: Marine heatwaves under global warming, Nature, 560, 360–364, 2018.

Garcia-Gorriz, E. and Garcia-Sanchez, J.: Prediction of sea surface temperatures in the western Mediterranean Sea by neural networks using satellite observations, Geophysical research letters, 34, 2007.

Garrabou, J., Coma, R., Bensoussan, N., Bally, M., Chevaldonné, P., Cigliano, M., Díaz, D., Harmelin, J.-G., Gambi, M. C., Kersting, D., et al.: Mass mortality in Northwestern Mediterranean rocky benthic communities: effects of the 2003 heat wave, Global change biology, 15, 1090–1103, 2009.

Garrabou, J., Gómez-Gras, D., Medrano, A., Cerrano, C., Ponti, M., Schlegel, R., Bensoussan, N., Turicchia, E., Sini, M., Gerovasileiou, V., et al.: Marine heatwaves drive recurrent mass mortalities in the Mediterranean Sea, Global Change Biology, 28, 5708–5725, 2022.

Giamalaki, K., Beaulieu, C., and Prochaska, J.: Assessing predictability of marine heatwaves with random forests, Geophysical Research Letters, 49, e2022GL099 069, 2022.

Giorgi, F.: Climate change hot-spots, Geophysical Research Letters, 33, 2006.

Guo, Y., Zhang, S., Yang, J., Yu, G., and Wang, Y.: Dual memory scale network for multi-step time series forecasting in thermal environment of aquaculture facility: A case study of recirculating aquaculture water temperature, Expert Systems with Applications, 208, 118 218, 2022.

Haghbin, M., Sharafati, A., Motta, D., Al-Ansari, N., and Noghani, M. H. M.: Applications of soft computing models for predicting sea surface temperature: a comprehensive review and assessment, Progress in earth and planetary science, 8, 1–19, 2021.

Han, M., Feng, Y., Zhao, X., Sun, C., Hong, F., and Liu, C.: A convolutional neural network using surface data to predict subsurface temperatures in the Pacific Ocean, IEEE Access, 7, 172 816–172 829, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C., Benthuysen, J. A., Burrows, M. T., Donat, M. G., Feng, M., et al.: A hierarchical approach to defining marine heatwaves, Progress in Oceanography, 141, 227–238, 2016.

Holbrook, N. J., Scannell, H. A., Sen Gupta, A., Benthuysen, J. A., Feng, M., Oliver, E. C., Alexander, L. V., Burrows, M. T., Donat, M. G., Hobday, A. J., et al.: A global assessment of marine heatwaves and their drivers, Nature Communications, 10, 1–13, 2019.

Holbrook, N. J., Sen Gupta, A., Oliver, E. C., Hobday, A. J., Benthuysen, J. A., Scannell, H. A., Smale, D. A., and Wernberg, T.: Keeping pace with marine heatwaves, Nature Reviews Earth & Environment, 1, 482–493, 2020.

Hornik, K.: Approximation capabilities of multilayer feedforward networks, Neural networks, 4, 251–257, 1991.

Hsieh, W. W.: Nonlinear canonical correlation analysis by neural networks, Neural Networks, 13, 1095–1105, 2000.

Jacox, M. G., Alexander, M. A., Siedlecki, S., Chen, K., Kwon, Y.-O., Brodie, S., Ortiz, I., Tommasi, D., Widlansky, M. J., Barrie, D., et al.: Seasonal-to-interannual prediction of North American coastal marine ecosystems: forecast methods, mechanisms of predictability, and priority developments, Progress in Oceanography, 183, 102 307, 2020.

Jacox, M. G., Alexander, M. A., Amaya, D., Becker, E., Bograd, S. J., Brodie, S., Hazen, E. L., Pozo Buil, M., and Tommasi, D.: Global seasonal forecasts of marine heatwaves, Nature, 604, 486–490, 2022.

Jacques-Dumas, V., Ragone, F., Borgnat, P., Abry, P., and Bouchet, F.: Deep learning-based extreme heatwave forecast, Frontiers in Climate, 4, 2022.

Juza, M., Fernández-Mora, À., and Tintoré, J.: Sub-Regional Marine Heat Waves in the Mediterranean Sea From Observations: Long-Term Surface Changes, Sub-Surface and Coastal Responses, Frontiers in Marine Science, 2022.

Lee, D., Won, K., Park, M., Choi, H., and Jung, S.: An analysis of mass mortalities in aquaculture fish farms on the southern coast in Korea, Ocean Policy Research, 33, 1–16, 2018.

Leroux, S., Brankart, J.-M., Albert, A., Brodeau, L., Molines, J.-M., Jamet, Q., Le Sommer, J., Penduff, T., and Brasseur, P.: Ensemble quantification of short-term predictability of the ocean dynamics at kilometric-scale resolution: A Western Mediterranean test-case, Ocean Science Discussions, 2022, 1–36, 2022.

Li, X., Liu, B., Zheng, G., Ren, Y., Zhang, S., Liu, Y., Gao, L., Liu, Y., Zhang, B., and Wang, F.: Deep-learning-based information mining from ocean remote-sensing imagery, National Science Review, 7, 1584–1605, 2020.

Liu, J., Zhang, T., Han, G., and Gou, Y.: TD-LSTM: Temporal dependence-based LSTM networks for marine temperature prediction, Sensors, 18, 3797, 2018.

Madec, G., Delecluse, P., Crepon, M., and Chartier, M.: A three-dimensional numerical study of deep-water formation in the northwestern Mediterranean Sea, Journal of Physical Oceanography, 21, 1349–1371, 1991.

Marbà, N., Jordà, G., Agusti, S., Girard, C., and Duarte, C. M.: Footprints of climate change on Mediterranean Sea biota, Frontiers in Marine Science, 2, 56, 2015.

McMillin, L. M.: Estimation of sea surface temperatures from two infrared window measurements with different absorption, Journal of geophysical research, 80, 5113–5117, 1975.

Merchant, C. J., Embury, O., Bulgin, C. E., Block, T., Corlett, G. K., Fiedler, E., Good, S. A., Mittaz, J., Rayner, N. A., Berry, D., et al.: Satellite-based time-series of sea-surface temperature since 1981 for climate applications, Scientific Data, 6, 1–18, 2019.

Pastor, F. and Khodayar, S.: Marine heat waves: Characterizing a major climate impact in the Mediterranean, Science of The Total Environment, p. 160621, 2022.

Pastor, F., Valiente, J. A., and Khodayar, S.: A warming Mediterranean: 38 years of increasing sea surface temperature, Remote sensing, 12, 2687, 2020.

Pinardi, N., Arneri, E., Crise, A., Ravaioli, M., and Zavatarelli, M.: The physical, sedimentary and ecological structure and variability of shelf areas in the Mediterranean sea (27), The sea, 14, 1243–330, 2006.

Rivetti, I., Fraschetti, S., Lionello, P., Zambianchi, E., and Boero, F.: Global warming and mass mortalities of benthic invertebrates in the Mediterranean Sea, PloS one, 9, e115 655, 2014.

Rodrigues, R. R., Taschetto, A. S., Sen Gupta, A., and Foltz, G. R.: Common cause for severe droughts in South America and marine heatwaves in the South Atlantic, Nature Geoscience, 12, 620–626, 2019.

Sarkar, P. P., Janardhan, P., and Roy, P.: Prediction of sea surface temperatures using deep learning neural networks, SN Applied Sciences, 2, 1458, 2020.

Schlegel, R. W., Oliver, E. C., and Chen, K.: Drivers of marine heatwaves in the Northwest Atlantic: The role of air–sea interaction during onset and decline, Frontiers in Marine Science, 8, 627 970, 2021.

Tang, B., Hsieh, W. W., Monahan, A. H., and Tangang, F. T.: Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial Pacific sea surface temperatures, Journal of Climate, 13, 287–293, 2000.

Tangang, F., Hsieh, W., and Tang, B.: Forecasting the equatorial Pacific sea surface temperatures by neural network models, Climate Dynamics, 13, 135–147, 1997.

Taylor, J. and Feng, M.: A deep learning model for forecasting global monthly mean sea surface temperature anomalies, Frontiers in Climate, 4, 178, 2022.

Tran, T. T. K., Bateni, S. M., Ki, S. J., and Vosoughifar, H.: A review of neural networks for air temperature forecasting, Water, 13, 1294, 2021.

Vogt, L., Burger, F. A., Griffies, S. M., and Frölicher, T. L.: Local drivers of marine heatwaves: a global analysis with an earth system model, Frontiers in climate, p. 49, 2022.

Wei, L. and Guan, L.: Seven-day Sea Surface Temperature Prediction using a 3DConv-LSTM model, Frontiers in Marine Science, p. 2606, 2022.

Wolff, S., O'Donncha, F., and Chen, B.: Statistical and machine learning ensemble modelling to forecast sea surface temperature, Journal of Marine Systems, 208, 103 347, 2020.

Wu, A., Hsieh, W. W., and Tang, B.: Neural network forecasts of the tropical Pacific sea surface temperatures, Neural networks, 19, 145–154, 2006.

Xiao, C., Chen, N., Hu, C., Wang, K., Gong, J., and Chen, Z.: Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach, Remote Sensing of Environment, 233, 111 358, 2019.

Xie, J., Zhang, J., Yu, J., and Xu, L.: An adaptive scale sea surface temperature predicting method based on deep learning with attention mechanism, IEEE Geoscience and Remote Sensing Letters, 17, 740–744, 2019.

Zanetta, F., Nerini, D., Beucler, T., and Liniger, M. A.: Physics-constrained deep learning postprocessing of temperature and humidity, Artificial Intelligence for the Earth Systems, 2023.

Figure legends and Table

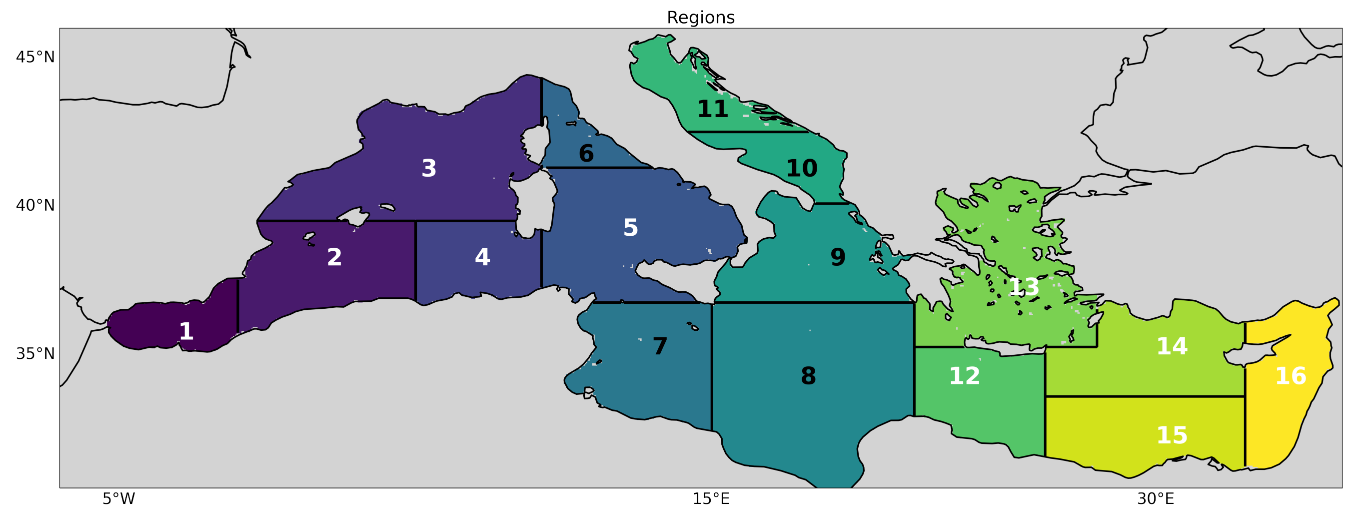

Figure 1. Mediterranean Sea regional subdivision and corresponding indices.

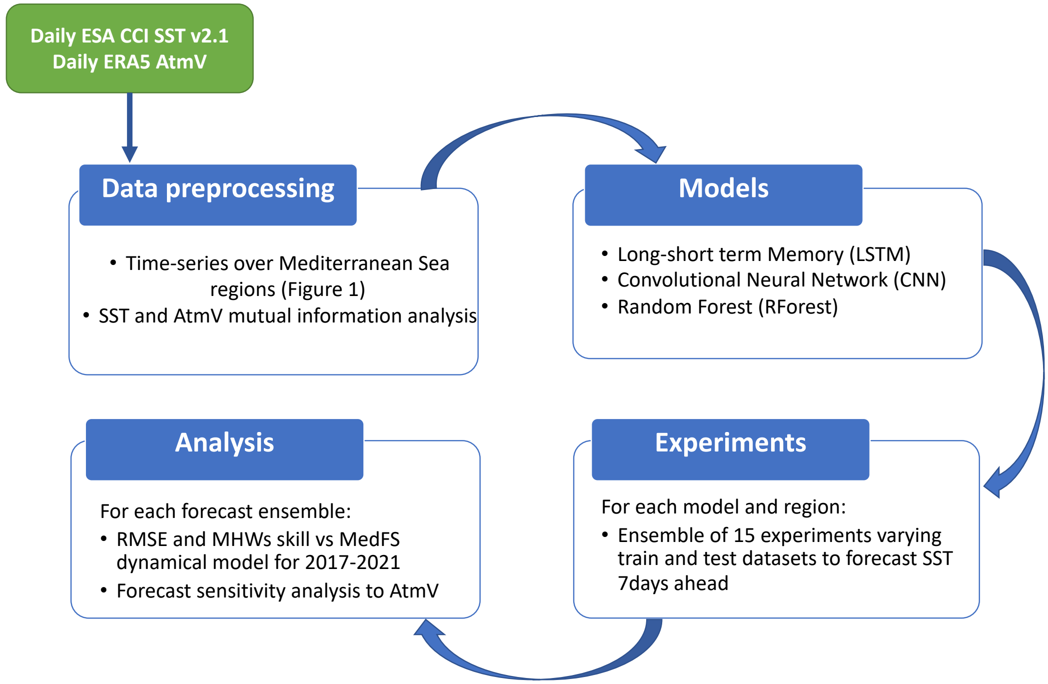

Figure 2. Flow diagram used in this study.

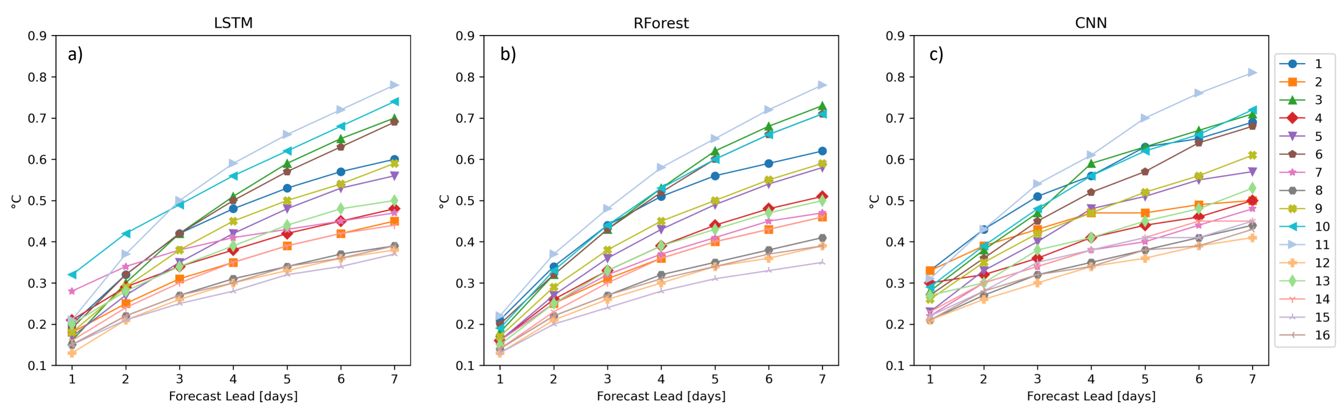

Figure 3. ML networks' performance for the SST daily predictions in terms of Root Mean Square Error (RMSE) in the Mediterranean Sea 16 regions (different colours) for (a) Long-short term memory networks, (b) Random Forest, (c) Convolutional Neural Networks.

	WIND		GEO		SLP		LAT		SENS		INC		MM	
Region ID	LAG0	LAG7	LAG0	LAG7	LAG0	LAG7	LAG0	LAG7	LAG0	LAG7	LAG0	LAG7	LAG0	LAG7
1	0.03	0.03	0.47	0.38	0.14	0.14	0.03	0.02	0.1	0.08	**0.64**	**0.55**	0.25	0.21
2	0.01	0.04	0.46	0.39	0.14	0.12	0.06	0.03	0.12	0.10	**0.71**	**0.63**	0.25	0.21
3	0.08	0.04	0.46	0.36	0.13	0.1	0.03	0.03	0.11	0.08	**0.75**	**0.62**	0.28	0.23
4	0.05	0.06	0.48	0.41	0.11	0.11	0.03	0.05	0.07	0.07	**0.68**	**0.60**	0.25	0.21
5	0.08	0.06	0.52	0.42	0.12	0.11	0.05	0.06	0.08	0.07	**0.79**	**0.68**	0.26	0.21
6	0.06	0.05	0.46	0.38	0.12	0.10	0.05	0.05	0.08	0.10	**0.62**	**0.75**	0.26	0.31
7	0.07	0.05	0.48	0.41	0.12	0.10	0.06	0.06	0.06	0.06	**0.81**	**0.73**	0.21	0.17
8	0.07	0.06	0.49	0.43	0.13	0.10	0.06	0.08	0.12	0.12	**0.87**	**0.79**	0.22	0.18
9	0.06	0.06	0.53	0.45	0.14	0.12	0.03	0.05	0.09	0.07	**0.78**	**0.67**	0.26	0.22
10	0.05	0.04	0.53	0.43	0.13	0.12	0.04	0.04	0.10	0.10	**0.73**	**0.62**	0.31	0.26
11	0.06	0.04	0.47	0.41	0.14	0.12	0.05	0.08	0.11	0.10	**0.79**	**0.66**	0.36	0.31
12	0.05	0.05	0.49	0.41	0.16	0.11	0.06	0.05	0.10	0.08	**0.82**	**0.73**	0.22	0.18
13	0.04	0.05	0.56	0.48	0.17	0.13	0.05	0.05	0.12	0.09	**0.77**	**0.64**	0.29	0.25
14	0.05	0.06	0.57	0.46	0.23	0.18	0.06	0.08	0.10	0.07	**0.83**	**0.72**	0.24	0.21
15	0.05	0.05	0.51	0.42	0.20	0.18	0.06	0.09	0.10	0.09	**0.84**	**0.74**	0.22	0.19
16	0.06	0.04	0.59	0.48	0.26	0.22	0.08	0.09	0.18	0.17	**0.85**	**0.74**	0.23	0.20

Table 1. Mutual Information between SST and AtmV variables not lagged in time (LAG0) and with 7 days lag (LAG7). The Lag-time is expressed in days. Bold values identify the highest Mutual Information values for LAG0 and LAG7

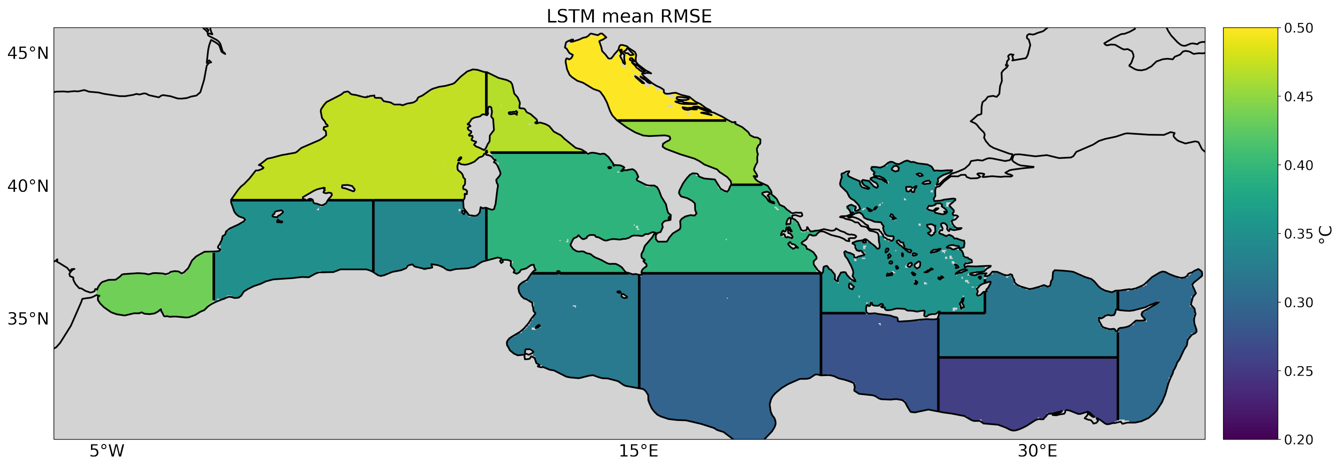

Figure 4. ~~a) ML networks performance for the SST daily predictions in terms of Root Mean Square Error (~~LSTM mean RMSE ~~) against the MedFS forecast system performance and b) variation of F1 score~~ for ~~MHWs occurrence prediction with the forecast lead time~~ REXP for ~~(left column) Western Mediterranean, (middle column) Central Mediterranean, (right column) Eastern Mediterranean. The sampling uncertainty of~~ each ~~prediction in (a) is illustrated by the whiskers. RMSE errors represented by the solid lines represent the reference experiments.~~ region

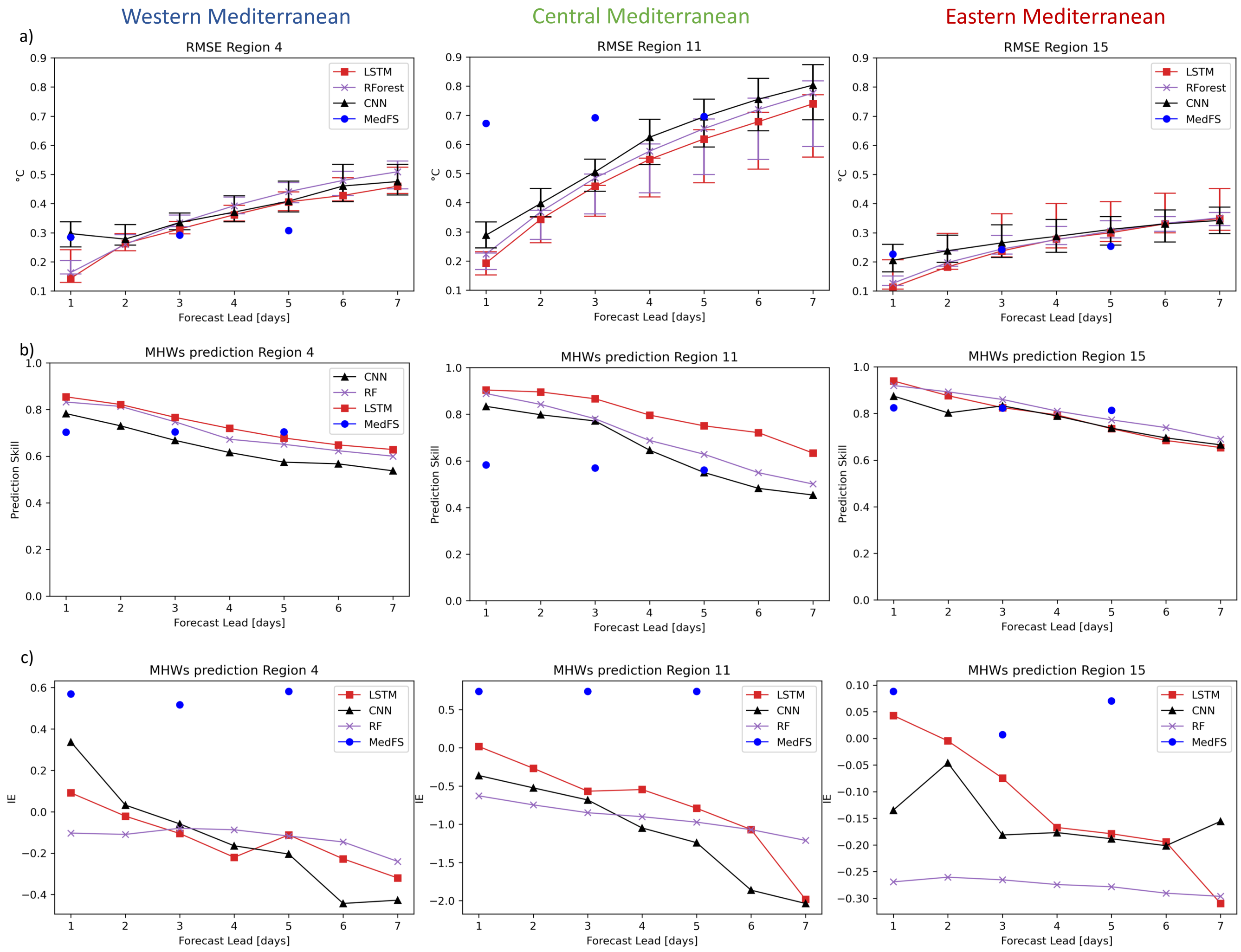

Figure 5. a) ML networks performance for the SST daily predictions in terms of Root Mean Square Error (RMSE) against the MedFS forecast system performance and b) variation of F1 score for MHWs occurrence prediction with the forecast lead time c) variation of forecasted MHWs intensity mean error (IE) with the forecast lead time for (left column) Western Mediterranean, (middle column) Central Mediterranean, (right column) Eastern Mediterranean. The sampling uncertainty of each prediction in (a) is illustrated by the bar. RMSE errors represented by the solid lines represent the reference experiments.

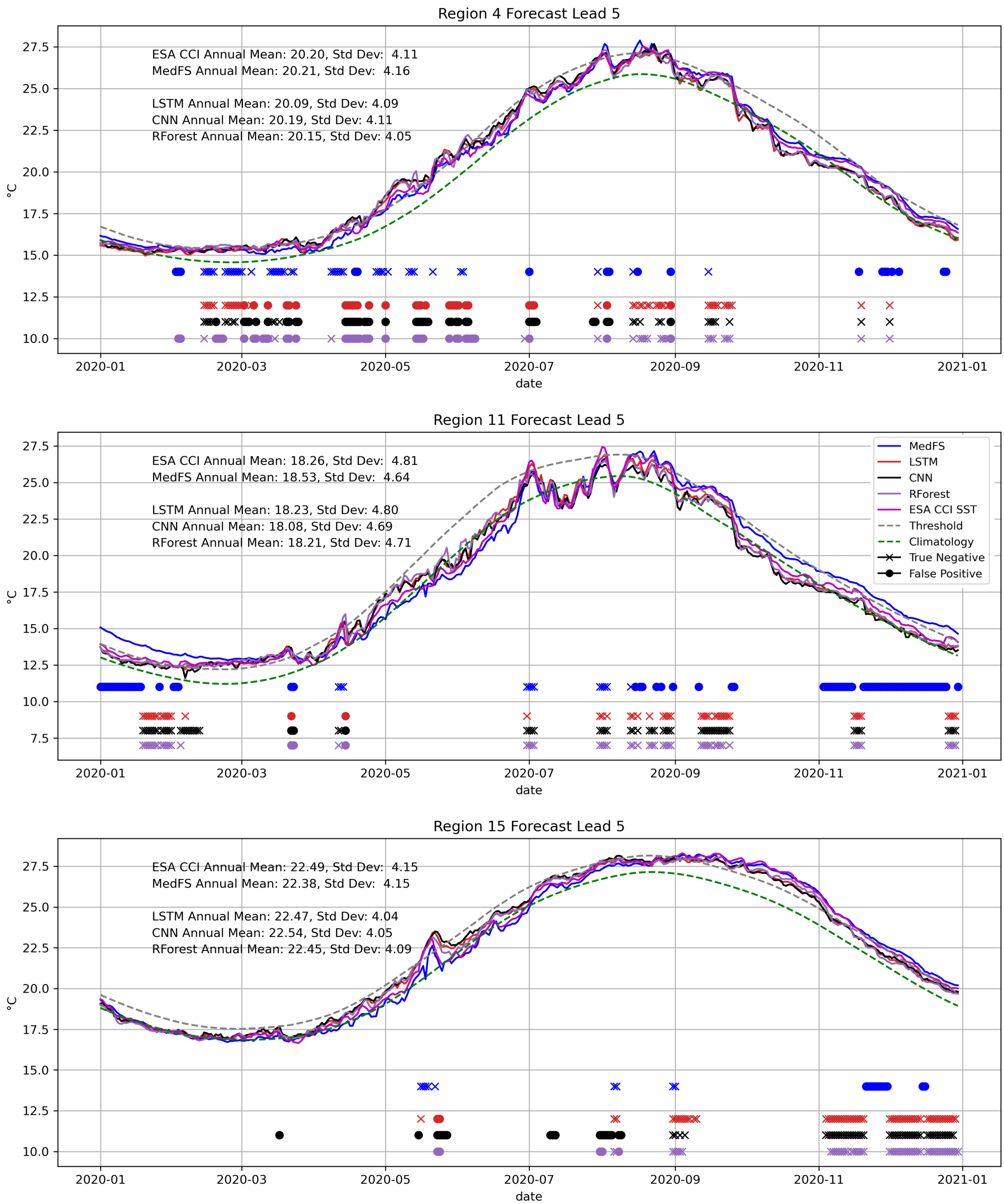

Figure 6. Time series of observed SST (ESA CCI SST) and predicted SST by the ML techniques (LSTM, CNN, RForest) and by MedFS at 5 days forecast lead time during 2020 for: (top panel) Western Mediterranean, (middle panel) Central Mediterranean and (bottom panel) Eastern Mediterranean. The 90th percentile threshold to define MHWs is represented in gray and the daily climatology in green. Crosses correspond to misses alarms (False Negative) and points to false alarms (False Positive) in the forecasts output in predicting MHWs. Colors refer to the different ML techniques.

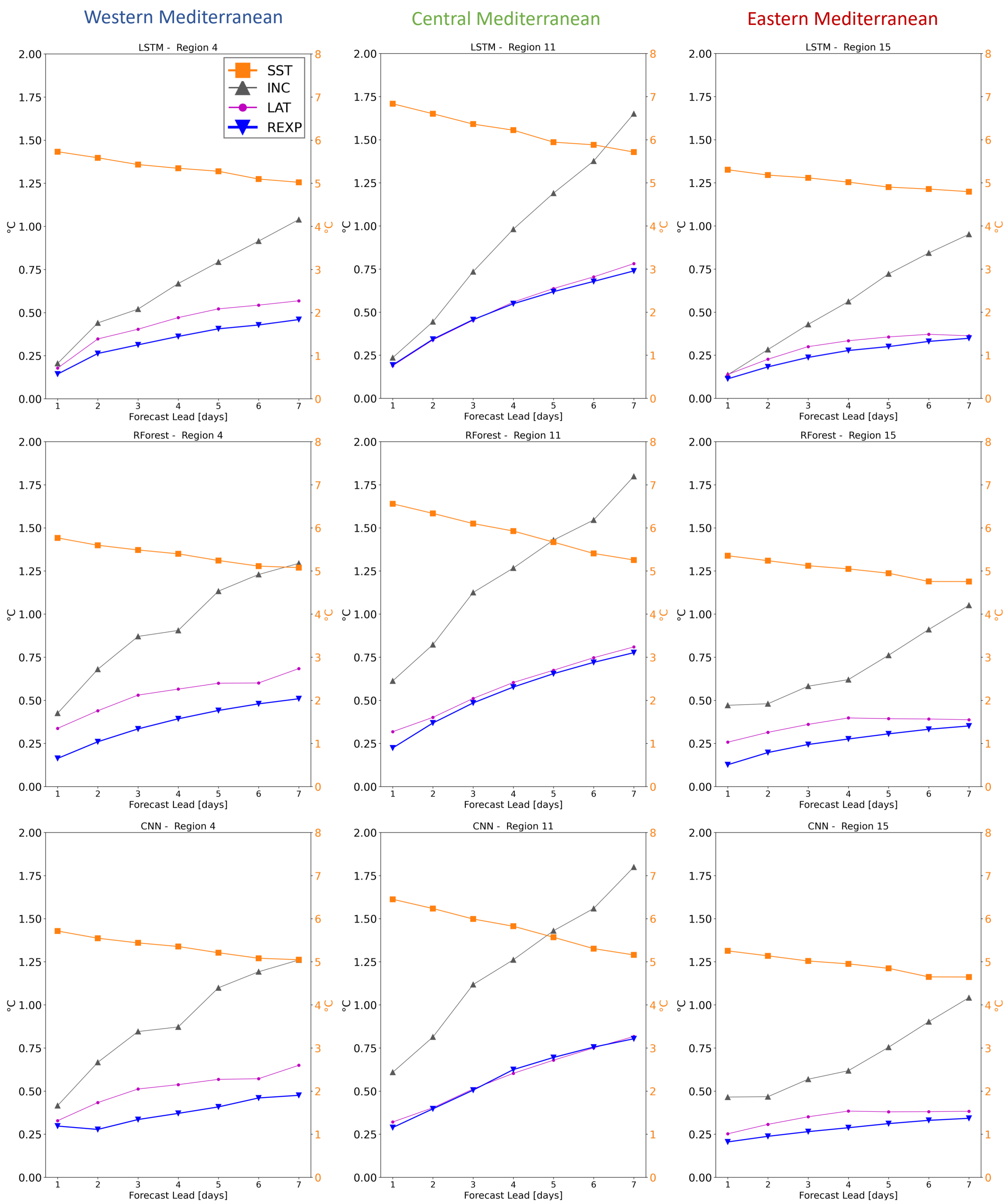

Figure 7. SST Root Mean Square Error (RMSE) of the sensitivity experiments *SEXPs* for each forecast lead time for: (left column) Western Mediterranean, (middle column) Central Mediterranean, (right column) Eastern Mediterranean. The labels indicate, for each experiment, the driver that has been shuffled. Right y axis (orange ticks) refers to SST driver experiment (orange solid line with "x" marker), left y axis indicates the other drivers error.

	MHW not predicted	MHW predicted
MHW not observed	TN	FP
MH predicted	FN	TP

Table 2. Accuracy in predicting MHWs

	MHW not predicted	MHW predicted
MHW not observed	TN	FP
MH predicted	FN	TP

Table 3. Accuracy in predicting MHWs

	LSTM		RForest		CNN	
Region ID	L1	L7	L1	L7	L1	L7
1	**0.17**	**0.59**	0.21	0.62	0.23	0.62
2	**0.15**	**0.45**	**0.15**	**0.45**	0.19	0.48
3	**0.16**	**0.70**	0.18	0.73	0.23	0.76
4	0.14	**0.45**	**0.16**	0.50	0.29	0.47
5	**0.13**	**0.58**	0.16	**0.58**	0.21	0.63
6	**0.19**	**0.65**	0.20	0.71	0.25	0.69
7	**0.14**	**0.43**	0.16	0.47	0.19	0.47
8	0.14	**0.39**	**0.13**	0.40	0.22	0.40
9	**0.14**	**0.57**	0.17	0.59	0.31	0.64
10	**0.16**	**0.63**	0.19	0.71	0.27	0.72
11	**0.19**	**0.74**	0.22	0.78	0.28	0.80
12	**0.12**	**0.38**	0.13	0.39	0.29	0.39
13	**0.13**	**0.50**	0.15	**0.50**	0.18	0.52
14	**0.12**	**0.44**	0.14	0.46	0.20	0.47
15	**0.11**	**0.35**	0.13	**0.35**	0.20	0.36
16	0.16	0.41	**0.14**	**0.38**	0.21	0.38

Table 4. ML network performance for the SST daily predictions in terms of Root Mean Square Error (RMSE) for Long-short term memory (LSTM), Random Forest (RForest) and Convolutional Neural Network (CNN) at the first day of forecast (Forecast Lead 1, L1) and the 7th days of forecast (Forecast Lead 7, L7) . Bold values identify the best performance (i.e. lowest RMSE) for L1 and L7.

		FPR			FNR			F1 Score		
Region ID	**Methods**	L1	L3	L5	L1	L3	L5	L1	L3	L5
4-WM	LSTM	0.04	0.04	0.07	0.13	0.33	0.39	**0.86**	**0.78**	0.71
	RForest	0.03	0.06	0.09	0.19	0.25	0.34	0.83	0.75	0.65
	CNN	0.07	0.04	0.05	0.17	0.42**	0.51**	0.78	0.67	0.57
	MedFS	0.07**	0.07**	0.06**	0.29**	0.31	0.31	0.70	0.70	**0.72**
11-CM	LSTM	0.02	0.02	0.02	0.1	0.17	0.38	**0.90**	**0.87**	**0.73**
	RForest	0.03	0.02	0.030	0.12	0.30	0.49	0.89	0.78	0.63
	CNN	0.06	0.03	0.01	0.12**	0.31**	0.59**	0.84	0.78	0.56
	MedFS	0.27**	0.27**	0.27**	0.17	0.19	0.2	0.6	0.58	0.58
15-EM	LSTM	0.03	0.05	0.1	0.06	0.21**	0.29**	**0.94**	0.84	0.74
	RForest	0.03	0.04	0.07	0.1	0.19	0.28	0.92	**0.86**	0.77
	CNN	0.15**	0.12**	0.15**	0.02	0.09	0.25	0.86	0.85	0.73
	MedFS	0.13	0.11	0.12**	0.12	0.15	0.16	0.82	0.82	**0.81**

Table 5. ML networks' performance for the MHWs daily predictions in terms of False Positive Rates (FPR), False Negative Rates (FNR) and F1 score for Long-short term memory (LSTM), Random Forest (RForest), Convolutional Neural Network (CNN) and the Mediterranean Forecasting System (MedFs) at the first day of forecast (Forecast Lead 1, L1), at the third days of forecast (Forecast Lead 3, L3) and at the 5th days of forecast (Forecast Lead 5, L5). Bold values identify the best F1 scores at L1, L3 and L5. Stars (**) identify the highest rates of FPR and FNR.