Marine Heatwaves in the Mediterranean Sea: A Convolutional Neural Network study for extreme event prediction

Antonios Parasyris[1], Vassiliki Metheniti[1], Nikolaos Kampanis[1] and Sofia Darmaraki[1]

[1]*Foundation for Research and Technology-Hellas, Institute of Applied and Computational Mathematics, 70013 Heraklion, Greece*

*Correspondence to: Antonios Parasyris (antonisparasyris@gmail.com)

Abstract: In recent decades, the Mediterranean Sea has experienced a notable rise in the occurrence and intensity of extreme warm temperature events, referred to as Marine Heatwaves (MHWs). Hence, the ability to forecast Mediterranean MHWs in the short term is an area of ongoing research. Here, we introduce a novel machine learning (ML) approach, specifically tailored for short-term predictions of MHWs in the basin, using an Attention U-Net Convolutional Neural Network. Trained on daily Sea Surface Temperature (SST) Aanomalies (SSTA) and gridded fields of MHW presence and absence between 1982- 2017, our model generates a spatiotemporal forecast of MHW occurrence up to 7 days in advance. To ensure robust performance, we explore various configurations, including different forecast horizons and U-net architectures, number of input days, features, and different subset splits of train-test datasets. Comparative analysis against a persistence benchmark reveals an improvement of 15% in forecasting accuracy of MHW presence for a 7-day forecast horizon. We also demonstrate an improvement of MHW prediction accuracy as the forecast horizon decreases, albeit with a smaller discrepancy between the persistence benchmark which also results in high accuracy for the 3-day forecasts. Our proposed ML methodology offers a data-driven prediction of MHWs with reduced computational requirements, which can be applied across different regions of the global ocean, providing relevant stakeholders and management authorities with essential lead time for implementing effective mitigation strategies.

Keywords: Marine Heatwaves; Mediterranean Sea; Machine learning; Convolutional Neural Networks; Extreme event forecasting; U-net, Attention

1 Introduction

Since the 1980s, the Mediterranean Sea has experienced a mean Sea Surface Temperature (SST) increase of approximately 0.041°C/year, which is twice the global average (Pisano et al., 2022). As a consequence, Mediterranean MHWs have increased in frequency, intensity and spatial coverage, with profound disruptions on marine ecosystems and communities that rely on them (Smith et al., 2021); Specifically, MHWs have caused numerous mass mortalities of native and the migration of invasive species in the Mediterranean Sea (Garrabou et al., 2022), threatening the region's rich marine biodiversity and commercially valuable fish stocks (Lacoue-Labarthe et al., 2016). Consequently, the ability to forecast MHWs has become central to the field of extreme oceanic events in the basin, as it enables the development of proactive measures for the mitigation of subsequent and potentially adverse effects on marine ecosystems and socio-economic activities.

Essential to the early prediction of MHWs are advanced monitoring systems and forecasting models, which have proven reliable on global (Schultz et al., 2021; Balaji, 2021) and regional applications (Konsta et al., 2023), in addition to sophisticated climate models that offer a comprehensive understanding of MHW drivers by simulating the complex interactions between atmospheric conditions and local oceanic circulation patterns (Darmaraki et al., 2019a). In the Mediterranean Sea, MHWs have been predicted in the short term by the regional Copernicus Mediterranean Forecasting System, which successfully captured phases of the summer 2022 event (Mcadam et al., 2023), while climate models have projected MHW frequency and characteristics throughout the 21st century, under different climate change scenarios (Darmaraki et al., 2019b; Konsta et al., 2023). However, forecasting systems and state-of-the-art numerical models present significant challenges due to the inherent uncertainties and chaotic nature of the climate system in addition to the associated high computational costs.

Thus, there is a growing interest in the use of Machine Learning (ML) techniques, particularly data-driven, Deep Learning models, to produce short- and long-term forecasts more efficiently (Schultz et al., 2021; Balaji, 2021). Deep Learning models are used as surrogate models that overcome the computational constraints present in classical numerical weather prediction (NWP) models. By training on large observational and/or model datasets, Deep Learning models can generate an extensive ensemble of forecasts, enhancing and complementing traditional numerical weather prediction NWP models (Chattopadhyay et al., 2020). Compared to traditional methods, ML models typically encounter fewer issues with bias (Jacox et al., 2022) and are especially skilled at capturing and representing intricate and nonlinear dynamics in data (Hornik, 1991).

Current research predominantly focuses on the forecast of SST (Taylor and Feng, 2022), which allows the establishment of contextualized thresholds based on specific requirements and conditions. This approach requires additional processing and expertise to identify MHWs (Hobday et al., 2016), which may complicate the application of these predictions by end users and management operators, a challenge that can be circumvented through data-driven spatiotemporal MHW prediction algorithms (Sun et al., 2023). Globally, the forecasting of SST using shallow ML techniques, such as linear regression and various statistical methods, is not new and dates back to the 1970s (Anding and Kauth, 1970; Fdez-Riverola et al., 2002; Mcmillin, 1975). During the last decade, the field has seen a significant shift towards the application of Deep Learning methods. These include Recurrent Neural Networks, long short-term memory networks (LSTM) (Xiao et al., 2019; Liu et al., 2018), Convolutional Neural Networks (CNN) (Han et al., 2019), and hybrid methodologies that combine these techniques (Taylor and Feng, 2022). These advancements have enhanced the accuracy and capabilities of SST predictions.

In contrast, direct forecasting of MHWs has received less attention. For instance, Giamalaki et al. (2022), employed a Random Forests method and successfully forecasted spatiotemporal occurrence of MHWs in the North Pacific Ocean. However, their approach exhibited limitations in accurately predicting the intensity and duration of these events, underscoring the challenges in developing robust predictive models for such complex phenomena (Schultz et al., 2021). More recently, Sun et al. (2023) advanced the field by training a hybrid model, that integrated CNNs with LSTM layers, to predict the occurrence and spatial development of MHWs. Their study proposed an innovative approach, combining CNN-LSTM architectures with a U-net CNN regression

model to forecast both SST anomalies and binary classification maps, indicative of the presence or absence of MHWs in the future. These outputs served as indicators of MHWs when they exceeded certain thresholds. This hybrid method focused on predicting the spatial distribution of MHWs rather than generating simple time-series forecasts and demonstrated significant potential for predicting MHWs with a lead time of up to 7 days. Such predictive capabilities can be particularly useful for the implementation of timely and effective mitigation strategies in regions highly vulnerable to MHW impacts, such as the Mediterranean Sea.

Nevertheless, the use of ML techniques for the early prediction of Mediterranean MHWs is an emerging field of research. One of the pioneering efforts has used aArtificial Nneural Nnetworks to predict the seasonal and inter-annual variability of SST in the western Mediterranean as well as the impact of the severe summer MHW of 2003 (Garcia-Gorriz and Garcia-Sanchez, 2007). By training on a variety of meteorological variables from 1960 to 2005, including 2-meter air temperature, wind, and sea level pressure, the study achieved reliable monthly SST predictions. More recently, Bonino et al. (2023a) evaluated the effectiveness of various ML algorithms, including LSTM networks, CNNs, and Random Forests on the forecast of daily SST, with a weekly lead time across 16 Mediterranean sub-basins. Their study focused on predicting SSTs that exceeded specific thresholds, indicating the potential occurrence of MHWs. This approach successfully reduced computational costs associated with the processing of 2D temperature fields, albeit at the expense of detailed information on the spatial variability of anomalous SST within each sub-basin. The results of this study outperformed outputs from the Copernicus Mediterranean Sea Analysis and Forecasting System, a sophisticated model that provides daily forecasts of ocean variables (Coppini et al., 2023). Overall, the application of ML techniques has showcased considerable potential in advancing our comprehension and forecasting of MHWs in the Mediterranean Sea.

Here, we combine well-established methodologies and tools to create a novel configuration of an Attention U-Net CNN, specifically tailored for the forecast of MHWs in the Mediterranean Sea. Using a minimal set of Regional Climate System Model (RCSM) variables, primarily the daily, spatial distribution of Sea Surface Temperature Anomalies (SSTA) SST (anomalies) and spatiotemporal information on the presence and absence of MHWs in the basin, produced by a MHW identification algorithm, we train the Attention U-Net CNN model to forecast MHW occurrence 3 and 7 days in advance. This work is structured as follows: Section 2 describes the ML technique, the input variables that were used to train the U-Net model and the error metrics used for validation. Results on MHW forecasting, based on various neural network configurations are shown in Section 3 followed by the Discussion and Conclusions on Section 4.

2 Methodology

2.1 Study area

The Mediterranean is a semi-enclosed, transitional area surrounded by the temperate zone to the north and the subtropical zone to the south and east. Its complex topography features sharp mountains, mild coastlines and desert regions, creating a region highly sensitive to climate change. The Mediterranean Sea is connected to the Atlantic Ocean through the Strait of Gibraltar to the west and to the Black Sea via the Bosphorus Strait to the northeast, which serves as the main freshwater inflow for the eastern basin. It has a mean depth of 1500m (Bethoux et al., 1999) and is divided in several sub-basins, including the Ionian, Tyrrhenian, Aegean, Adriatic, Alboran, Balearic, and Ligurian Seas (Fig. 1). The region is also characterized by a distinctive thermohaline circulation driven by surface heat and water losses. This unique circulation system balances the excess evaporation over the Mediterranean, contributing to a net buoyancy flux towards the atmosphere, which plays a crucial role in the regional climate dynamics.
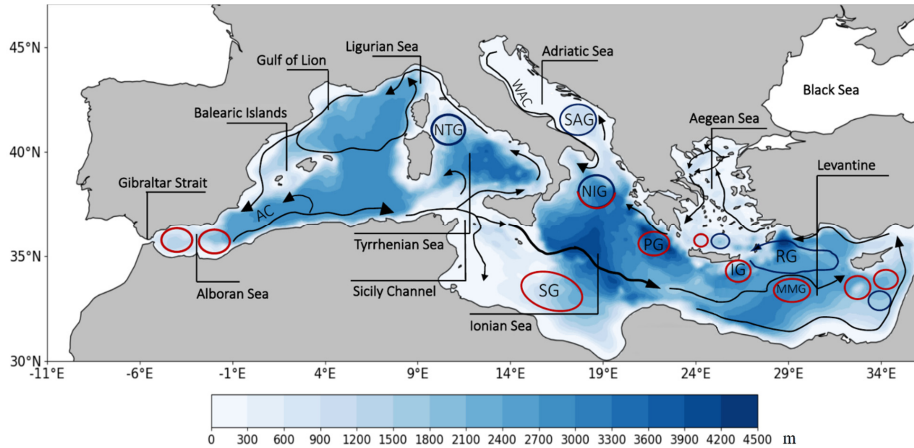
Figure 1: *The bathymetry, main circulation and sub-basins of the Mediterranean Sea. Key features of the surface circulation are shown with black arrows, with anticyclonic and cyclonic systems represented by red and blue circles, respectively. Acronyms: AC, Algerian Current; NTG, North Tyrrhenian Gyre; SG, Sidra Gyre; WAC, Western Adriatic Current; SAG, Southern Adriatic Gyre; NIG, Northern Ionian Gyre; PG, Pelops Gyre; IG, Ierapetra Gyre; RG, Rhodos Gyre; MMG, Mersa-Matruh Gyre. Bathymetry (m) is given in colors based on the CNRM-RCSM6 model. The figure is reprocessed based on Menna et al. (2022), Velaoras et al. (2024) and Darmaraki et al. (2024)*

2.2 Input Data

To identify surface MHWs in the Mediterranean Sea and train the ML algorithm for their short-term prediction, we obtain daily gridded, SST outputs from the fully-coupled, regional climate system model CNRM-RCSM6, that run on hindcast mode between 1982-2017 (Sevault, 2024; Darmaraki et al., 2019a). The model (NEMOMED12) covers the entire Mediterranean Sea domain, has a 6-8 km horizontal resolution, with a varying vertical resolution, over 75 vertical levels in the ocean ~~(NEMOMED12)~~ (Beuvier et al., 2012; Waldman et al., 2017) and its lateral boundary conditions come from ERA-Interim (Berrisford et al., 2009).

130 The ML method is trained, tested and validated on a combination of 2 types of input variables: (1) daily, gridded ~~SST anomalies (~~SSTA~~)~~ relative to the 1982-2017 period and (2) gridded fields of MHW presence/absence. The daily spatial coverage of surface MHW is computed using the updated version of the MHW detection algorithm by Hobday et al. (2016), available at https://github.com/coecms/xmhw (Petrelli, 2022). According to this definition, a MHW occurs when the local SST is above a 30-year climatological period (1982-2014) and a
135 threshold of the 90th percentile of SST, for at least 5 consecutive days. The MHW identification method yields gridded, binary classification masks of daily MHW presence/absence (1/0) for the entire Mediterranean Sea domain, between 1982-2017. The predominant classification of grid points as MHW-absent (0) for most days of the year results in an imbalanced input dataset that affects the forecasting accuracy of any data-driven model (Bonino et al., 2023a; Sun et al., 2023).

140

To achieve a more balanced input dataset and reduce memory requirements during training the gridded MHW occurrence fields are first downsampled. This process considers a non-overlapping 2x2 submatrix around a point and assigns label (1) in the presence of at least one MHW-affected pixel, or a label of (0) in its absence, leading to a decreased spatial resolution of MHW occurrence fields. The same approach is followed for points in the
145 immediate proximity to the coast. The final spatial resolution of the Mediterranean Sea resolves to 128x272 points, which reduces the spatial resolution of the model to approximately 12-16 km. Despite the reduction in the spatial resolution, this approach increases the proportion of MHW presence to 14% (from the initial 7.7%) across the entire domain, while still preserving a significant portion of the complicated Mediterranean Sea features and the geographical location of each MHW. To match the spatial resolution of the gridded MHW occurrence fields,
150 we also downsample the daily SSTA of the Mediterranean Sea, by means of spatially averaging within a non-overlapping 2x2 submatrix around a point. Daily SSTA are further normalized to a range [0,1], aligning with the scale of the gridded MHW occurrence data, with a view to improve the prediction accuracy of the U-Net CNN (Xiao et al., 2019).

155 As an input to the CNN model for a given day, we insert the gridded fields of MHW occurrence and SSTA from preceding days, targeting the prediction of MHW spatial coverage 3 and 7 days ahead (Fig. 2). The limited set of input variables is selected as an effective approach to balance the risk of overfitting with computational efficiency in our method. This decision was further informed by a lagged correlation analysis, which revealed moderate correlations between SSTA and atmospheric variables such as air temperature, latent heat flux and shortwave
160 radiation, with SSTA displaying the highest lagged autocorrelation (not shown). We then perform sensitivity tests to assess the impact of varying the number of preceding days as input, specifically 0, 2 and 4 days for the MHW occurrence fields and 5, 10 and 14 days for SSTA, on the model's predictive ability. ~~A c~~Cosine and sine functions are also inserted as input features, indicating the yearly seasonal cycle, assuming a unique combination of values for each day of the year, spanning the period from 1/1/1982 to 31/12/2017:

$$\cos_t = \cos\left(\frac{2\pi t}{365}\right); \forall t \in \{1, 2, \ldots, Ndays\}$$

$$\sin_t = \sin\left(\frac{2\pi t}{365}\right); \forall t \in \{1, 2, \ldots, Ndays\} \ , \hspace{3cm} \text{(Eq. 1)}$$

where the total number of days is denoted with *Ndays.*

2.3 The Attention U-net CNN model

In this study, we employ a data-driven approach, which means that the nature, quantity and quality of our data (e.g. number of available years, temporal discontinuities in data, spatiotemporal resolution and forecasted variable characteristics including class imbalance, seasonality, randomness etc.) dictate the methodological choices for forecasting MHW occurrence across different time horizons. Using a neural network architecture, enhanced with attention mechanisms (Attention U-Net CNN mode) which emphasize key features while suppressing noise and capturing important spatiotemporal patterns, we focus on 3-day and 7-day forecast horizons. This is deemed a sufficient time for proactive decision-making by authorities and local stakeholders (Giamalaki et al., 2022) before a MHW incident (Giamalaki et al., 2022).

2.3.1 Attention U-net Architecture

Here, the prediction of MHWs is considered a supervised classification/regression challenge, for which we employ the specific neural network architecture illustrated in Fig. 2, based on the U-net CNN architecture proposed by Ronneberger et al. (2015). Due to the use of several intermediate (hidden) layers, the method is
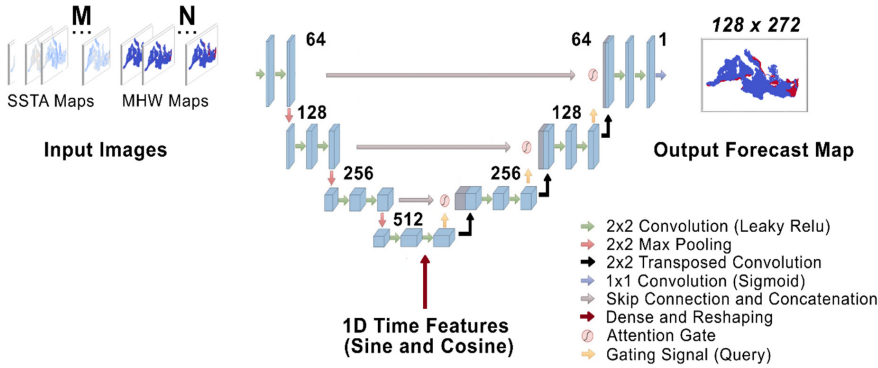
Figure 2*: Attention U-net CNN Model architecture to forecast MHW presence/absence maps. N and M are the number of input frames containing spatiotemporal information on MHW presence/absence and SST anomaly, respectively. Each map corresponds to daily frequency of input data and has a matrix size of 128 x 272. –U-net figure is adapted by Ibtehaz and Rahman (2020)*

classified as a deep neural network category, with the architecture including both a contracting and expanding path. The contracting path is responsible for extracting features from the input data and progressively reducing the size of the feature maps through downsampling, which, in our case, consists of a series of three Max pooling layers (https://www.tensorflow.org/api_docs/python/tf/keras/layers/MaxPooling2D last accessed online 3/2024).
185 This process effectively increases the model's ability to perceive broader spatial relationships. As the contracting path progresses, the number of channels in the feature maps is doubled to improve feature capture across different scales. Conversely, the expanding path uses up-sampling operations to restore the feature maps to their original dimensions, consisting of 3 deconvolution layers, using a Conv2D Transpose (https://www.tensorflow.org/api_docs/python/tf/keras/layers/Conv2DTranspose last accessed online 3/2024) for
190 the decoding path, as shown in Fig. 2.

Moreover, the model incorporates skip connections to integrate both local and global features, enabling the network to utilize information from various scales simultaneously. Attention gates are placed before concatenating the skip connections, automatically learning to focus on target structures of varying shapes and
195 sizes. This design is expected to improve the model's prediction accuracy while maintaining minimal computational overhead (Oktay et al., 2018). The architecture also includes 2D Convolutional layers with 2x2 kernels between all layers. The number of neurons in each layer follows an exponential pattern with a base of 2, increasing in the contracting path and decreasing in the expanding path. The sine/cosine time features from Eq. 1 are incorporated at the bottom of the encoding path, after passing through dense and reshaping layers to obtain
200 the same spatial dimension as the layer with which are concatenated.

Due to its versatility, this method has been previously used in studies of image segmentation, pattern identification (Oktay et al., 2018; Srivastava et al., 2014) spatiotemporal forecasting (Jacques-Dumas et al., 2022), and downscaling at higher resolutions with minimal computational cost (Doury et al., 2023) and has been
205 shown to significantly enhance the accuracy of forecasts. The goal of this network is to determine the probability of each gridpoint being classified as either MHW-present (1) or MHW-absent (0) for predictions of 3 or 7 days ahead, thus forecasting the spatiotemporal probability of MHW occurrence.

2.3.2 U-net CNN model Training

A common practice in a neural network approach is the splitting of a dataset into the training, testing and
210 validation subsets. Here, the validation dataset consists of the early years from 1982 to mid-1986, the training dataset comprises the years from mid-1986 to mid-2013 and the final years selected for testing and validation of the model span from mid-2013 to 2017. Following the split, each of the three subsets (train, test and validation) undergoes random internal shuffling to reduce memorization effects and increase the robustness of the forecasting tool. The tendency for a model to memorize rather than learn meaningful patterns -known as
215 overfitting- is a critical challenge in neural network training, as it leads to excessive tuning to the training data, and causes the model to capture noise and irrelevant details, ultimately compromising its ability to generalize effectively to new, unseen data. To further reduce overfitting effects, we employ a "dropout" approach, which involves random deactivation of a specified number of nodes, set to 30% here, during each training step

(Srivastava et al., 2014). An early stopping/best saving checkpoint mechanism is also implemented, whereby the training process halts if there is no improvement for a predetermined number of epochs and a selected validation metric (i.e., mean squared error, accuracy, recall, f1-score etc.).

Throughout the training of our model, we employ the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba, 2014), an optimization algorithm, which is based on two gradient descent methodologies, with a batch size of 4 and an initial learning rate of 0.001, set up to reduce by half on a plateau of 10 epochs, with a minimum learning rate of 10^{-4} and maximum number of 100 epochs. Evaluated across all the training samples, the Adam optimizer was used here to minimize the loss function, a key parameter of the Attention U-net CNN, which measures the discrepancy between predicted and actual values. The model's ability to successfully perform a given task is determined by the choice of the loss function and the effective reduction of prediction errors. Given the binary form of the MHW presence (1) and absence (0) fields, the use of a binary cross-entropy loss function is a common choice (Jacques-Dumas et al., 2022). In the case of extreme events and imbalanced datasets, where one of the two classes is underrepresented, the focal binary crossentropy is preferred (https://www.tensorflow.org/api_docs/python/tf/keras/losses/BinaryFocalCrossentropy last accessed online 3/2024) (Lin et al., 2020). This function further improves the effects of the standard binary cross-entropy by integrating two additional parameters, designed to reduce the influence of correctly classified samples and emphasize the importance of misclassified ones. We use this loss function, which in equation form reads as:

$$Focal\left(p_t\right)=-\sum^{N\cdot M}\alpha_t\left(1-p_t\right)^\gamma \log\left(p_t\right) \text{ where } p_t=\begin{cases} p & ,y=1 \\ 1-p & ,otherwise \end{cases}, \qquad \text{(Eq. 2)}$$

where y is the ground truth class, $p \in [0,1]$ is the predicted probability and ~~a~~α_t and γ are the tuning parameters, which are selected as 0.25 and 2 respectively, following Nguyen and Thai (2023). This adjustment leads to elevated loss values when misclassifications occur, steering the training process toward lower local minima of the loss function. Due to the non-linear nature of the patterns the neural network is trained on and despite the attempts to minimize the loss function, its convergence often leads to near local minima, impeding its performance. To improve the model's performance and ensure that minima values of the loss function maintain satisfactory accuracy levels in their prediction, several iterations with varying parametric choices are conducted on test cases (see Results Sect. 3.2, Table 1).

An additional aspect of neural network architecture is the selection of activation functions, which are applied to the outputs of each intermediate hidden layer. These functions have a key role in determining the operations applied to the input neurons and thereby the model's ability to generate an output (Sharma et al., 2020). In this study we overcome the limitations associated with the use of a standard Rectified Linear Unit (ReLU) activation function in handling negative input values, by using a Leaky ReLU version in all the intermediate layers (Maas et al., 2013). The Leaky ReLU function allows a small, non-zero gradient for negative inputs, effectively mitigating the vanishing ReLU problem, where neurons become inactive during training. This effect proves advantageous for non-linear prediction tasks, enabling the model to capture a broader range of input variations. The output (final) layer is equipped with a sigmoid activation function, which is appropriate for binary classification tasks, as it produces values within the range of [0,1]. These values express the probability for a grid point being affected by a MHW, with the classification threshold determined by the specific characteristics of the physical problem in

each instance. The process by which we select the appropriate classification threshold is further discussed in Sect. 3.1.

260 The U-net CNN model described here was implemented using the Keras API, using Tensorflow 2.9.2 in Python. The architecture outlined in Fig. 2 required 32 million parameters to train, and it run in parallel on 8 Nvidia A100 GPUs with a memory of 40 GB each, in one node of the p4d.24xlarge Amazon EC2 server. Each test case required approximately 4 hours to complete the training of 100 epochs, and the inference speed required seconds to calculate forecasts for each sample on the same server. We note that once the computationally expensive 265 training ends, the model can be deployed to any system to generate forecasts, with minimal computational overhead, given the appropriate input data. For instance, the inference time for the entire testing dataset, consisting of 1300 daily forecasts, required approximately 60 seconds on the aforementioned hardware configuration.

270 **2.4 Evaluation Metrics**

The model's performance is evaluated on a testing dataset consisting of samples that were obscured by the U-net CNN during its training phase. At this stage, we assess the model's ability to accurately predict unseen data and determine the optimal probability threshold, above which each grid point is categorized as a MHW-affected case. Throughout the training and validation process, standard metrics such as Recall, accuracy and the selected loss 275 function are calculated during each epoch for the training dataset and at the end of each epoch for the validation dataset (https://www.tensorflow.org/api_docs/python/tf/keras/metrics last accessed online 3/2024). The primary metric to evaluate the prediction skill of the Attention U-net CNN on future MHW occurrence is the True Positive Rate (TPR). This rate assesses the proportion of accurately predicted MHW occurrences relative to the total number of actual occurrences (Sun et al., 2023). The formula for calculating TPR is given by:

$$TPR = \frac{TP}{TP + FN},$$

280 (Eq. 3)

where TP represents True Positive predictions and FN denotes False Negative ones. In other words, the TPR metric quantifies the model's ability to accurately detect true MHW occurrence during testing. At the end of each epoch during the training process, validation Recall, a metric similar to TPR, is computed on the validation subset and serves as an early-stopping mechanism, a technique useful for both computational efficiency and preventing 285 overfitting. In contrast, the True Negative Rate (TNR) evaluates the model's ability to accurately predict the negative class labels, specifically the absence of a MHW, relative to the total number of non-MHW occurrences. This metric reads as:

$$TNR = \frac{TN}{TN + FP},$$

(Eq. 4)

where TN represents the True Negatives and FP the false positives. Combining these two rates, the Forecast 290 Accuracy Rate (FAR) composite metric can be used to assess the model's overall forecasting ability. Unlike the TPR and TNR which focus on a single class, the FAR considers both the correct and incorrect MHW predictions across all classes. In formula form, FAR is defined as:

$$FAR = \frac{TN + TP}{TN + FP + TP + FN},$$
(Eq. 5)

providing a percentage-wise estimation of forecast accuracy.

295

To obtain an overall TPR, TNR and FAR as single numerical indicators for each sample in the testing dataset we separately average the metrics defined in (3), (4) and (5) over the Mediterranean Sea domain. This approach provides a single, spatially independent numerical value for each error metric corresponding to each sample in the testing dataset. The final TPR, TNR and FAR for each forecast horizon are determined by averaging each of these metrics across all test samples.

300

2.5 Persistence Benchmark

A standard approach to evaluating the predictive skill of our ML model is to compare its output forecast with a climatological baseline derived from a Persistence benchmark model, which assumes that MHW presence or absence, within the next 3 or 7 days, remains constant throughout the forecast period. Specifically, the Persistence benchmark uses MHW presence/absence fields from 3 or 7 days prior to an event as both the input and the forecast for the target MHW conditions. By operating under the assumption that recent MHW conditions persist into the forecast period, the Persistence model provides a baseline performance level against which we assess whether the U-net CNN model adds predictive value beyond simple temporal persistence. Despite the simplicity of this assumption, using the Persistence benchmark model as a reference dataset is a meaningful approach for short-term forecasting (Parasyris et al., 2022), given the relatively slow changes in SST, which lead to minimal variations in MHW presence on most consecutive days. Improvements demonstrated by the output forecast of a neural network model that surpasses the basic Persistence model are essential for accurately forecasting the onset and dissipation of MHWs.

305

310

315

3 Results

3.1 Assessing the probability of MHW occurrence

The output probabilities of MHW occurrence are converted into binary classification masks as an initial step to assess the forecast ability of our ML model. In particular, the U-net CNN generates maps where each pixel value ranges between 0 and 1, representing the probability of each grid point being a negative (MHW-absent) or positive (MHW-present) class (referred to as forecast probability). A range of forecast probability thresholds (hereafter, threshold) are then evaluated from 0.05 to 0.9, in increments of 0.05, as a trial-and-error technique, to find the one which maximizes a specific accuracy metric (TPR, TNR, FAR). A prediction based on each threshold is subsequently generated for the entire testing dataset and compared with the true MHW occurrences originally identified using the climate model output. Following the work of Sun et al. (2023), we first compute the spatially averaged TPR and TNR separately for each threshold, and further average their sum for all samples within the test set. The optimal thresholds for the 3-day (Fig. 3a) and 7-day (Fig. 3b) forecast scenarios are then determined by maximizing this Combined Mean (CM) of TPR and TNR, which is formulated as:

320

325

330

$$CM = \frac{TPR + TNR}{2},$$ (Eq. 6)

In both scenarios, increasing the threshold results in a higher TNR and lower TPR, as more predictions are classified as MHWs absences, the accuracy of which is thereby improved. The intersection point in each forecast scenario (Fig. 3, red star) represents an equilibrium between sensitivity and specificity in classification terms.

335 This point maximizes the CM (Fig. 3, green lines), optimizing predictive performance of both the positive and
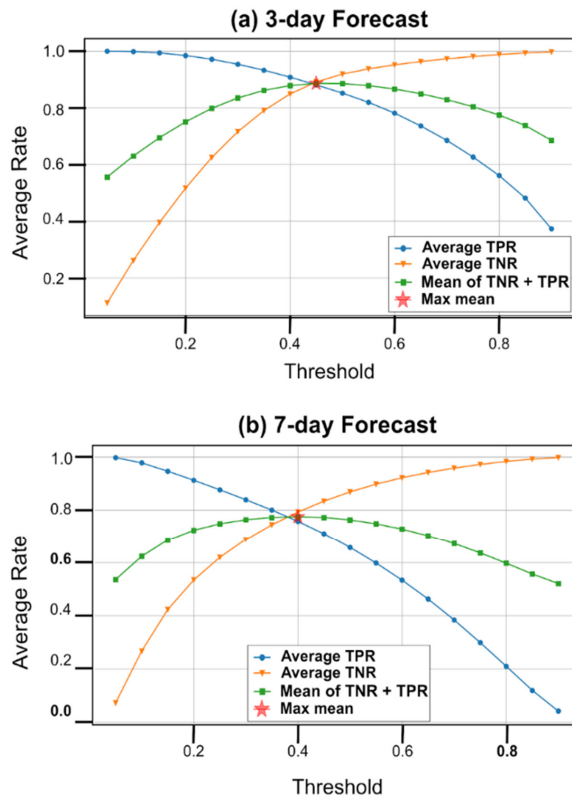
Figure 3: *Threshold selection for the a) 3-day and (b) 7-day forecast scenario based on the maximization of the combined mean (green line) of TPR (blue line) and TNR (orange line). The threshold increments range from 0.05 to 0.9 and the maximal point is indicated by a red star.*

negative MHW occurrences. Specifically, the optimal threshold for the 3-day forecast is identified at 0.45, yielding a maximum CM of 0.89 (Fig. 3a), whereas the 7-day forecast achieves a CM of 0.77, at an optimal threshold of 0.4 (Fig. 3b). At a high threshold of 0.85, both scenarios exhibit a TNR close to 1, indicating no false negatives, as most predictions are classified as negative. However, at this threshold, the respective TPR is 0.48 for the 3-day forecast and 0 for the 7-day forecast (Fig. 3, blue lines). This suggests that for the 3-day forecast, half of the MHWs were correctly classified despite the excessively high threshold. In contrast, the 7-day forecast shows a decline in performance, as TPR is approximately zero (Fig. 3b).

In the context of optimizing the threshold selection, TPR (Eq. 3) and TNR (Eq. 4) are prioritized over FAR (Eq. 5), as the primary goal is to accurately predict MHW occurrences (True Positives) rather than minimizing false alarms (FP, FN). Given that FAR includes all grid points in the domain (128 x 272), it relies on a larger denominator, resulting in lower overall values. This causes the FAR curve to shift toward higher threshold values, consequently reducing the number of True Positives, which is suboptimal. Nevertheless, FAR is useful as a cumulative metric, and is later used to assess the spatial overall performance of the optimal model configuration.

3.2 Sensitivity Analysis on Input Variables

To further optimize the configuration of our U-net CNN, we perform a sensitivity analysis on the input variables employed in training the model. Specifically, 9 distinct experiments are examined for both the 3-day and 7-day forecast scenarios, in which we varied the number of SSTA days and MHW presence/absence window preceding the target prediction date. The number of experiments was primarily limited by the memory constraints of our computational setup during training. Additionally, we prioritized simplicity to ensure minimal requirements during forecasting, enhancing the tool's practicality as a viable alternative for MHW prediction. For each experiment we evaluate the corresponding TPR, TNR, CM and FAR metrics and determine the best-performing setup of our model based on the maximum CM. The decision to prioritize CM (over FAR) reflects the greater influence of TPR on CM, aligning with our primary objective of accurately predicting positive MHW occurrences.

In the case of the 3-day forecast scenario, experiment 5 achieves the highest CM, incorporating M=10 preceding timesteps of SSTA and N=2 preceding timesteps of MHW presence/absence maps as input variables, with a total of 31.926.629 parameters trained (Table 1). Although this experiment demonstrates the highest TPR (0.881) and the second highest TNR (0.891) among all experiments, it does not yield the maximum FAR, which is observed in experiments 3 and 2. Overall, the TPR and TNR variations across most of the sensitivity tests in the 3-day forecast scenario, remain within a 5% range, indicating a relative stability of the model's performance and robustness to small changes in input variables. However, a marked deterioration in performance is observed in experiments 1 and 4, with a TPR declining by up to 16%, compared to other experiments. This is due to the SSTA being the sole input variable (N=0) of this configuration.

Table 1*: Sensitivity experiments on the input variables of the U-net CNN model and the associated prediction metrics for the 3-day forecast scenario. The input variables configuration for each experiment is indicated with the number of preceding timesteps of SSTA (N) and the number of preceding timesteps (M) of MHW presence/absence maps. The highest evaluation metrics are highlighted in bold, for clarity. The TPR, TNR, CM and FAR represent True Positive Rate, True Negative Rate, Combined Mean and Forecast Accuracy Rate, respectively.*

Experiment	M	N	TPR	TNR	CM	FAR
1	5	0	0.785	0.726	0.756	0.783
2	5	2	0.844	0.898	0.871	**0.918**
3	5	4	0.864	0.890	0.880	**0.918**
4	10	0	0.758	0.759	0.758	0.786
5	10	2	**0.881**	0.891	**0.886**	0.900
6	10	4	0.858	**0.905**	0.882	0.902
7	14	2	0.865	0.893	0.879	0.917
8	14	4	0.867	0.850	0.859	0.914
9	Persistence		0.844	0.890	0.867	0.881

In comparison, the 7-day forecast produces slightly reduced metrics, indicating decreased accuracy over the extended forecast period (Table 2). In particular, the highest CM is identified in experiment 7, which incorporates M=14 preceding timesteps of SSTA and N=2 preceding timesteps of MHW presence/absence maps as input variables. While this experiment demonstrates a balanced performance, achieving a TPR of 0.757 and the second highest TNR (0.792), experiment 9 (the Persistence benchmark) exhibits the highest TNR (0.872) and a significantly lower TPR (0.653).

Compared to the 3-day forecast, the TPR and TNR variations across the sensitivity tests of the 7-day forecast, reach up to 14% (Table 2). For instance, TPR ranges from 0.653 in experiment 9 to 0.774 in experiment 4, while TNR ranges from 0.648 in experiment 4 to 0.872 in experiment 9. When excluding cases that lack MHW presence/absence maps as an input variable (N=0), the results show the same model stability (Table 1), ~~demonstrated in Table 1,~~ with TPR and TNR variations confined within a 5% range. Indeed, a marked deterioration in performance is observed in experiments 1 and 4, with TPR values ranging between 0.753-0.774 and TNR values between 0.688-0.648, respectively. This indicates that the model has limited ability to effectively ~~leverage~~draw upon patterns from past events ~~is limited~~ when information about MHW presence/absence maps is excluded.

415

Table 2 *As in Table 1 but for the 7-day forecast scenario.*

Experiment	M	N	TPR	TNR	CM	FAR
1	5	0	0.753	0.688	0.721	0.718
2	5	2	0.704	0.812	0.758	0.823
3	5	4	0.761	0.735	0.748	0.778
4	10	0	0.774	0.648	0.711	0.716
5	10	2	0.735	0.794	0.764	0.801
6	10	4	0.7565	0.759	0.757	0.788
7	14	2	**0.757**	0.792	**0.775**	0.802
8	14	4	0.756	0.754	0.755	0.780
9	Persistence		0.653	**0.872**	0.763	**0.829**

3.3 Performance of optimized U-net CNN configuration

420 In the following, we assess the forecasting ability of our U-net CNN model based on the best performing configurations of each forecast horizon. Specifically, we assess the predictive capability of experiment 5 for the 3-day forecast scenario and experiment 7 for the 7-day forecast scenario, which were determined to yield the highest performance metrics.

425 **3.3.1 Forecast Rates**

The performance of the binary MHW classification is evaluated across the entire testing dataset (mid 2013 to 2017), irrespective of temporal or spatial variations, using a confusion matrix. This tool quantifies the percentage of correctly and incorrectly classified grid points within the MHW spatial domain throughout all the samples of each forecast scenario, providing a comprehensive evaluation of the model's predictive accuracy and robustness.

430 The classification metrics used include the TPR and TNR as well as the False Positive Rate (FPR) for negative cases misclassified as positive and the False Negative Rate (FNR) for positive cases misclassified as negative, which are the complementary metrics to the TPR and TNR, respectively.

The model achieves high accuracy in correctly classifying both the MHW-affected (True Positive) and the non-

435 MHW grid points (True Negative) in the 3-day forecast scenario, with success rates of 88.1% and 89.1%, respectively (Fig. 4a). Incorrect predictions account for only up to 12% of the samples. In comparison, relatively lower rates of TPR (75.7%) and TNR (79.2%) are seen in the 7-day forecast (Fig. 4b), likely due to the reduced autocorrelation of the 7-day lagged input maps, compared to the 3-day case, with the proportion of misclassified cases increasing to 24%.

440

To complement the analysis of aggregate metrics, we further examine the predictive accuracy of our U-net CNN on a sample-by-sample basis within the testing dataset, spanning mid-2013 to 2017 (Fig. 5). Specifically, we compare the total number of predicted positive MHW grid points to the observed positives for each sample. By examining the alignment of data points with the line of parity (Y=X), we assess not only the overall predictive

445 skill but also patterns of systematic overprediction or underprediction within the dataset. This approach enables us to identify deviations from perfect agreement and provides insights into potential biases in the forecasts. For the 3-day forecast, a strong agreement is observed between the predicted and observed MHW occurrences, as 99.31% of the points fall within a tolerance of 3,500 grid points from the Y = X line (Fig. 5a). This tolerance threshold was empirically determined to optimize performance in the 3-day forecast scenario, enabling

450 meaningful comparisons with the 7-day forecast. Given that in geospatial analysis it is common to use a threshold based on a percentage of the total dataset to ensure stability and scalability (Xu et al., 2024), we chose a tolerance that represents approximately 10% of the total grid points in each map. In contrast, the 7-day forecast scenario shows reduced alignment, with 86.68% of samples meeting the same tolerance threshold. Furthermore, most points lie above the Y = X line, indicating a potential overestimation of MHW presence in this scenario (Fig. 5b).

455 These results highlight the ML model's diminished predictive reliability at extended forecast horizons, consistent with the inherent trade-offs in balancing sensitivity and specificity.
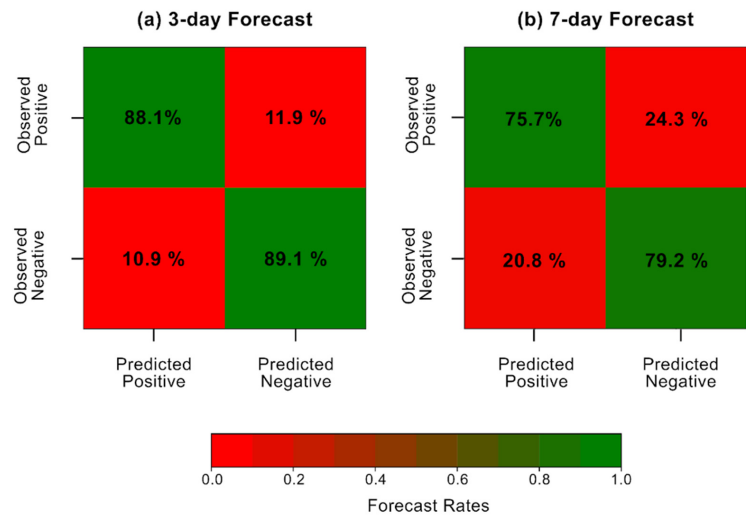
Figure 4: *Confusion plot for the (a) 3-day and (ii) 7-day forecast, showing the TPR (Observed Positive-Predicted Positive), TNR (Observed Negative-Predicted Negative), FPR (Observed Negative-Predicted Positive) and FNR (Observed Positive-Predicted Negative). Forecast rates are denoted in percentages within the boxes, taking into account all the samples of the testing dataset spanning mid-2013 to 2017.*
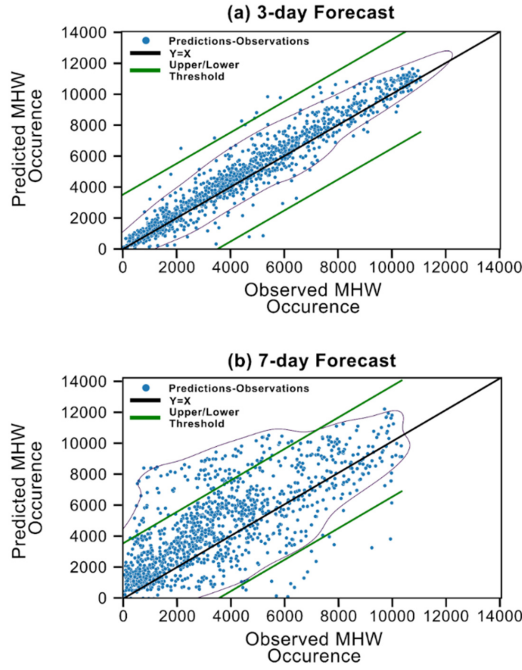
Figure 5: *Scatter plots of the total number of observed (x-axis) vs predicted (y-axis) points classified as MHW present (blue dots) per sample, for the a) 3-day and b) 7-day forecast throughout the entire testing dataset spanning mid-2013 to 2017. The upper and lower thresholds of 3500 points (green lines) are introduced for comparison purposes (see text).*

3.3.2 Comparison with the Persistence Benchmark

The final phase of evaluating the U-Net CNN configuration focuses on comparing its performance to the
460 Persistence benchmark model, which predicts MHW presence/absence based on lagged correlations over 3-day and 7-day intervals (Fig. 6). The Persistence model predicts MHW occurrence based on past observations, assuming that the most recent conditions are the best predictor of future events. The "noisy" TPR fluctuations observed across the test dataset can arise from rapid changes in the MHW presence/absence maps, which challenge the predictive stability of both models. This variability is further compounded in the U-Net CNN by the
465 inherent stochasticity in neural network predictions. However, in terms of the average TPR, the U-net CNN consistently outperforms the Persistence benchmark across both forecast scenarios (Fig. 6). For the 3-day forecast, the Persistence model achieves an average TPR of 0.844 (Fig. 6a), which declines to 0.653 in the 7-day

forecast (Fig. 6b), reflecting stronger performance in the shorter forecast horizons. In comparison, the reduction in TPR between the 3-day and 7-day forecast is less pronounced for the U-net CNN, declining from 0.881 to 0.757. This indicates that the U-net CNN exhibits a comparatively higher stability in maintaining predictive performance over longer forecast horizons. Overall, the best-performing experiments of the U-net CNN outperform the Persistence benchmark model, achieving higher values across all evaluation metrics in the 3-day forecast and in some select metrics for the 7-day forecast. This reflects the model's robustness in shorter forecast horizons and its ability to maintain competitive performance despite the challenges posed by longer forecast horizons.
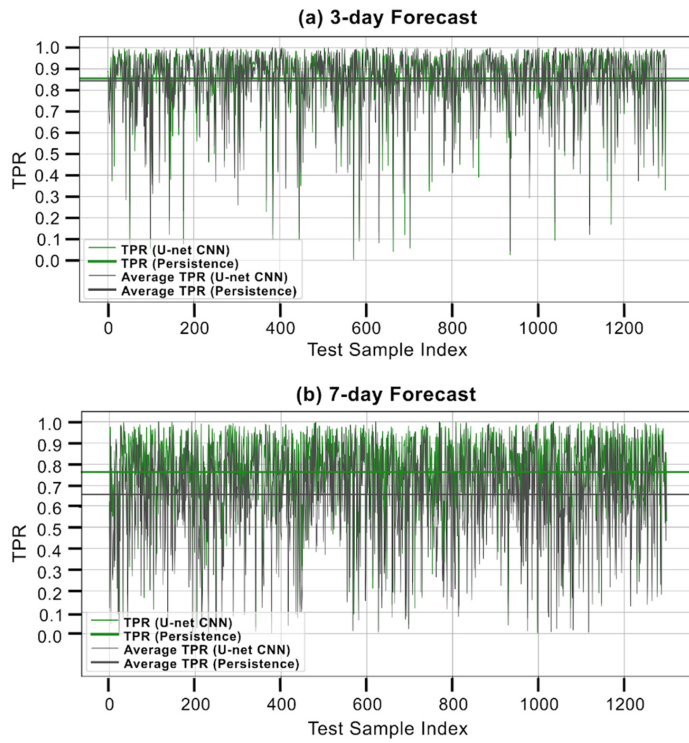
Figure 6: *TPR values of each sample across the entire 3.5-year test dataset spanning mid-2013 to 2017 for the best-performing experiment and Persistence benchmark of the a) 3-day and b) 7-day forecast. The TPR values of each U-net model are indicated in green lines whereas TPR of the Persistence model is shown in black. The averaged value across the entire 3.5-year test dataset is also shown with the same colors as a bold line, indicating our U-net method to be better overall.*

3.3.3 Evaluation of Spatial Prediction Accuracy

This section investigates the regions of the Mediterranean Sea displaying higher and lower susceptibility to prediction errors in our method. To this end, we examine the spatial distribution of the averaged FAR across all samples of the testing dataset (mid-2013 to 2017) for both forecast scenarios. This metric provides a cumulative assessment of the U-net CNN model's predictive performance, enhancing our understanding of regional prediction reliability, while guiding model refinements to improve forecast precision in identified areas of weakness.

Overall, the spatial distribution of FAR reveals significant differences between the 3-day and 7-day forecast scenarios. In the 3-day forecast, FAR values exceed 90% in the northwest Mediterranean, the Adriatic and Ionian Sea as well as the southeast Mediterranean basin, as opposed to the Balearic Islands, Alboran and Aegean Sea, where slightly lower FAR (80% - 90%) is displayed (Fig. 7a). In the 7-day forecast, FAR values are generally lower across the entire Mediterranean Sea. Specifically, the FAR values in the Aegean Sea, the coastal areas of the northern Mediterranean basin and he Alboran Sea, range between 60%-70%, with only the Ionian and Tyrrhenian Sea as well as the Levantine basin displaying FAR values approximately between 80%-92% (Fig. 7b).
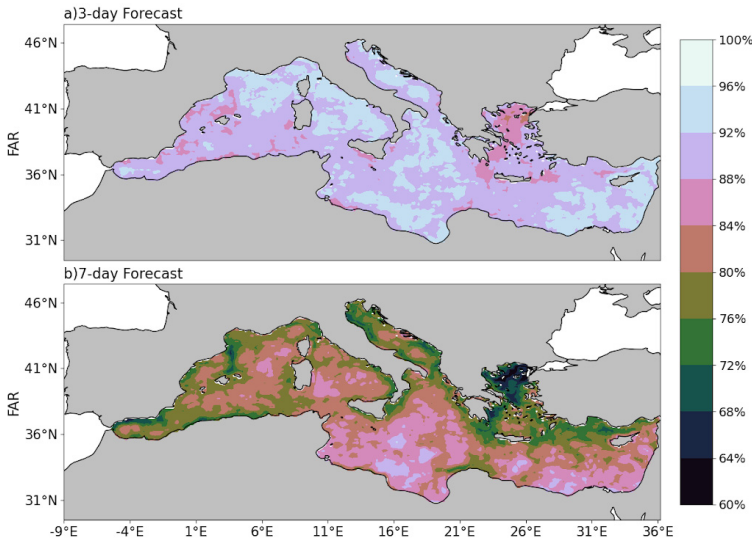
Figure 7: *Spatial distribution of the Forecast Accuracy Rate (FAR) in the Mediterranean Sea for the a) 3-day and b) 7-day forecast scenario, averaged across all samples of the testing dataset spanning mid-2013 to 2017.*

4 Discussion and Conclusions

This study presents the application of an Attention U-Net CNN to predict the spatiotemporal evolution of MHWs
in the Mediterranean Sea. The proposed model integrates an attention mechanism with a standard U-Net
architecture, combined with a focal binary cross-entropy loss function, among other key parameters. The model is
trained on SSTA ~~anomalies~~ and information on daily MHW presence/absence between 1982-2017 to predict
future MHW occurrence within 3-day and 7-day forecast horizons. Extensive hyper parameter tuning is carried
out to ensure the model's stability and performance within acceptable limits, alongside the implementation of a
specific threshold selection technique. The results, presented for both forecast horizons, highlight a decline in
forecasting skill as the prediction horizon increases, with our U-net CNN consistently outperforming the
Persistence benchmark model.

For the optimal thresholding technique of MHW probability, we prioritize the accurate detection of true MHW
instances (TPR) over minimizing the number of non-MHW instances (TNR). Specifically, we focus on the
maximization of the combined mean of TPR and TNR (CM), to ensure a balance between sensitivity and
specificity (Sun et al., 2023). As the threshold increases, a trade-off emerges between the TPR and TNR (green
line, Fig. 3), whose curves form a concave shape with a distinct maximum point. In cases where this maximum is
not well-defined, an optimal threshold can be determined either empirically (Fawcett, 2006), based on the
intersection of TPR and TNR (Fig. 3), or by maximizing the FAR metric (Sun et al., 2023).

The optimal configuration of the U-net CNN is then determined through sensitivity analysis, assessing the impact
of various input variable combinations on model performance. The best-performing configuration incorporates
M=10 preceding SSTA timesteps and N=2 preceding timesteps of MHW presence/absence for the 3-day forecast
(Table 1) and M=10 and N=4 for the 7-day forecast (Table 2). In both forecast scenarios, configurations relying
solely on SSTA data demonstrate the weakest performance, highlighting the importance of integrating both
temperature anomalies and prior MHW occurrences to achieve more reliable forecasting of these events.
Particularly for the 7-day forecast the variations in the CM and FAR metrics across the different experiments are
less pronounced when excluding experiments with a single input variable. In line with the findings of Sun et al.
(2023), this sensitivity analysis aims to optimize the model's performance on unseen data, while maintaining
simplicity for future end-users and balancing model complexity and generalization.

However, the assessment of the U-net CNN's forecast ability reveals distinct differences in predictive accuracy
between the short-term (3-day) and longer-term (7-day) forecast horizon. While the 3-day forecast achieves the
highest average TPR and TNR metrics overall, its discrepancy relative to the Persistence benchmark is modest,
potentially due to the high autocorrelation of the SSTA in both the benchmark and the 3-day forecast (Fig. 6a). In
contrast the 7-day forecast exhibits lower TNR and TPR values, with the performance gap between the
benchmark and the U-net CNN being more pronounced (Fig. 6b). Thus, our results indicate an improved
accuracy in shorter forecast horizons, in agreement with Bonino et al. (2023a). Despite demonstrating a decline in
TPR ~~between~~ from the 3-day to the 7-day ~~the two~~ forecast scenario~~s~~, the strength of the U-net CNN lies in

maintaining a higher TPR value than the Persistence benchmark in both forecast scenarios, suggesting an improved ability to predict MHWs across different temporal scales.

Based on the FAR metric, an improved forecasting performance (high FAR values) of the U-net CNN is revealed in the northwest Mediterranean basin, the Ionian Sea as well as the Levantine basin, as opposed to the lower FAR values observed in the Aegean and Alboran Sea, the Balearic Islands and northern coastal areas (7-day forecast). In the case of the Aegean Sea, the low FAR values may reflect challenges in detecting MHWs due to rapid SST fluctuations influenced by the Black Sea. These fluctuations can lead to swift onset and dissipation of MHWs (Mavropoulou et al., 2016), complicating their detection by ML methods. Additionally, the downsampling applied to address the high computational demands of our method (see Sect. 2.2) resulted in a reduced spatial resolution of the data. This reduction likely affected the model's performance in regions with complex topography, such as the Aegean Sea and coastal areas. This is likely due to the reduction in spatial resolution (halved to 12–16 km) near the coast, which results in the averaging of values and the loss of high-resolution information in these regions. While we acknowledge this limitation, we prioritize reliable forecasting of MHWs across the entire Mediterranean Sea, understanding that higher-resolution models, though more accurate in these areas, also come with increased computational demands. This is also observed in Bonino et al. (2023a), where the authors also report a reduced forecast ability of their neural network model around the Adriatic Sea, Balearic islands, the Alboran and Aegean Sea. Given that ocean circulation along the coast is primarily driven by local winds and can be influenced by offshore currents near complex topographic features, the exclusion of winds as an input dataset may have further reduced the forecast accuracy of our ML model in shallow coastal areas, where SST variations become more complicated (Berthou et al., 2024; Liu et al., 2025). However, Bonino et al. (2023a) found a low correlation between the wind speed and SST in all the Mediterranean sub-basins they considered. To do that, prior to the ML method, the authors calculated the Mutual Information index using SST and several variables of interest, indicating that the amount of information that can be gained from one variable by observing the other is low. While incorporating atmospheric variables into the training process could thus potentially enhance the model's ability to capture broader climatic influences on MHW occurrence, ultimately, we selected a limited set of training variables, in order to enhance the model's simplicity and replicability across both the training and prediction phases, following Sun et al. (2023).

It is important to note that the results of this study are based on the "straight split" methodology, where the early years of the dataset are used for training and the final years for testing. Given that ML models trained on recent data often achieve higher accuracy, due to the influence from recent climate patterns (recency effect; Lam et al., 2023), we have also explored the impact of reversing the training and testing dataset order. In particular, we carried out a sensitivity test, defined as the "opposite split", using the years 2013-2017 for training and 1982–1986 for testing, to assess whether our model's predictive skill is dataset-dependent or driven by climate change-induced temperature (see Supplementary Material Table S1). For the 7-day forecast horizon, we find improved TNR and TPR metrics when more recent data are inserted in the training dataset. This aligns with the findings of Lam et al. (2023), where recent climate trends, including the increased frequency of MHWs, were shown to enhance model effectiveness by mitigating~~frequency of MHWs, can enhance model effectiveness by mitigating~~ issues of data scarcity and imbalance in the training datasets. However, for the 3-day forecast, we find a

deterioration of both metrics, compared to the outputs of the "straight split" methodology (see Supplementary Material Table S1). This may be due to the longer duration of recent MHWs, which are less frequent in earlier years (Oliver et al., 2018), and thus may be poorly represented in the training dataset. In this study, we thus used the "straight split" methodology, as the training of a neural network with historical data to forecast information in the future reflects a more realistic approach.

Overall, the proposed U-net CNN model offers a computationally efficient alternative to traditional regional forecasting models for predicting MHWs in the Mediterranean Sea. Once trained, our approach maintains high spatiotemporal resolution while requiring minimal computational resources. The results indicate that the model performs better in the 3-day forecast compared to the 7-day forecast across all evaluation metrics and relative to the Persistence benchmark. This outcome is expected, given the higher autocorrelation of SSTA over shorter timeframes, consistent with other studies e.g.,such as Sun et al. (2023); Taylor and Feng (2022). Although the 7-day forecast holds greater practical value for the early prediction of extreme events, enabling more effective mitigation strategies, it remains a challenging task for ML methods, as reflected in the lower Persistence benchmark values over longer time horizons. In comparison, the 3-day forecast achieves higher overall metrics due to the shorter forecasting window and the inherently higher Persistence benchmark values. Nevertheless, the U-Net model still outperforms the benchmark in the 3-day forecast, albeit by a smaller margin than in the 7-day forecast. Given the model's success in predicting MHWs using a minimum input of variables in a region with diverse thermohaline and circulation patterns, such as the Mediterranean Sea (Benincasa et al., 2024), it is reasonable to assume that our methodology can be generalized to other basins and case studies. While a similar ML approach has demonstrated comparable forecasting performance in areas with similar data availability (Sun et al., 2023), factors such as grid size and computational resources may also influence the training process of the ML model. Notably, the primary challenge in applying the U-Net CNN for predicting MHWs in this study was achieving a balance between predictive accuracy and computational efficiency, an important consideration for the application of all ML methods.

As global warming accelerates, the increasing frequency and severity of MHWs pose significant challenges to marine ecosystems. Efficient and timely forecasting of MHWs is essential for effective marine management, enabling governments, industries, and coastal communities to take proactive measures, such as imposing temporary fishing bans, enhancing monitoring of vulnerable species, establishing marine protected areas, and launching public awareness campaigns to promote sustainable practices. ML-based approaches show promise for improving predictions of these events, particularly through architectures like the Attention U-Net CNN employed in this study, which reconstruct spatially distributed variables and generate high resolution predictions, within seconds to minutes, depending on available computational resources. This rapid forecasting ability is particularly advantageous for short-term predictions, as climate models typically require days of runtime and data assimilation to achieve comparable accuracy in operational settings (Coppini et al., 2023). For instance, a significant computational benefit was demonstrated using a neural network-based regional climate model (RCM) emulator that was trained to reproduce complex spatial structure and variability of near-surface temperature, simulated by an RCM (Doury et al., 2023). By learning the relationship between the low-resolution predictors and high-resolution surface variables over the RCM domain, this approach enables the low-cost generation of

high-resolution RCM simulation ensembles, which are useful for exploring local-scale uncertainties in present and future climate.

As computational capabilities advance, ensemble models, such as those used in time series regression (Bertsimas and Boussioux, 2023) could also improve CNN-based forecasting of MHWs, especially for long-term model predictions that are typically hindered by error propagation. Future research should consider training ML models on observed or remotely sensed data, despite their limitations (Abdelmajeed and Juszczak, 2024), as well as hybrid approaches, that combine the physical consistency of traditional models with the speed and adaptability of ML methods (Bonino et al., 2023b). Given that the accuracy and reliability of many ML models is compromised by their violation of fundamental physical principles (Chen et al., 2023), future efforts should also focus on addressing this limitation, by incorporating physical laws (Desai and Strachan, 2021) or analytical equations (Zanetta et al., 2023), approaches that, though in early stages, show promise.

Data availability: Data used are cited when firstly introduced and are available from corresponding authors upon request.

Author Contributions. Conceptualization: **A.P.** Data curation: **A.P., S.D.** Formal analysis: **A.P., S.D., V.M.** Funding acquisition: **S.D.** Investigation: **A.P., S.D.** Methodology: **A.P.** Project administration: **S.D., A.P.** Resources: **A.P., S.D., N.K.** Software: **A.P., V.M.** Supervision: **S.D.** Validation: **A.P., V.M.** Visualization: **A.P., S.D., V.M.** Writing – original draft preparation: **A.P., S.D.** Writing – review & editing: **S.D., V.M., A.P.**

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements: Computational time is granted by Meluxina under ID u100862 and AWS EC2 server under GRNET-EDYTE Amazon grant. We would also like to thank Florence Sevault and Samuel Somot for providing us with the CNRM-RCSM6 simulation and for their support in accessing the dataset. The simulations used in this work contributes to the Med-CORDEX initiative (www.medcordex.eu).

Financial support. Hellenic Foundation for Research and Innovation (H.F.R.I.) under the 3rd Call of "Research Projects to Support Post-Doctoral Researchers" scheme (Project Number 07077, acronym TExMed)

References

Abdelmajeed, A. Y. and Juszczak, R.: Challenges and Limitations of Remote Sensing Applications in Northern Peatlands: Present and Future Prospects, 10.3390/rs16030591, 2024.

Anding, D. and Kauth, R.: Estimation of sea surface temperature from space, Remote Sensing of Environment, 1, 217-220, https://doi.org/10.1016/S0034-4257(70)80002-5, 1970.

Balaji, V.: Climbing down Charney's ladder: machine learning and the post-Dennard era of computational climate science, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200085, 10.1098/rsta.2020.0085, 2021.

Benincasa, R., Liguori, G., Pinardi, N., and von Storch, H.: Internal and forced ocean variability in the Mediterranean Sea, Ocean Sci., 20, 1003-1012, 10.5194/os-20-1003-2024, 2024.

Berrisford, P., Dee, D., Fielding, K., Fuentes, M., Kallberg, P., Kobayashi, S., and Uppala, S. S.: The ERA-interim archive, 2009.

Berthou, S., Renshaw, R., Smyth, T., Tinker, J., Grist, J. P., Wihsgott, J. U., Jones, S., Inall, M., Nolan, G., Berx, B., Arnold, A., Blunn, L. P., Castillo, J. M., Cotterill, D., Daly, E., Dow, G., Gómez, B., Fraser-Leonhardt, V., Hirschi, J. J. M., Lewis, H. W., Mahmood, S., and Worsfold, M.: Exceptional atmospheric conditions in June 2023 generated

a northwest European marine heatwave which contributed to breaking land temperature records, Communications Earth & Environment, 5, 287, 10.1038/s43247-024-01413-8, 2024.

655 Bertsimas, D. and Boussioux, L. J. a. p. a.: Ensemble modeling for time series forecasting: an adaptive robust optimization approach, 2023.

Bethoux, J. P., Gentili, B., Morin, P., Nicolas, E., Pierre, C., and Ruiz-Pino, D.: The Mediterranean Sea: a miniature ocean for climatic and environmental studies and a key for the climatic functioning of the North Atlantic, Progress in Oceanography, 44, 131-146, https://doi.org/10.1016/S0079-6611(99)00023-3, 1999.

660 Beuvier, J., Béranger, K., Lebeaupin Brossier, C., Somot, S., Sevault, F., Drillet, Y., Bourdallé-Badie, R., Ferry, N., and Lyard, F.: Spreading of the Western Mediterranean Deep Water after winter 2005: Time scales and deep cyclone transport, Journal of Geophysical Research: Oceans, 117, https://doi.org/10.1029/2011JC007679, 2012.

Bonino, G., Galimberti, G., Masina, S., McAdam, R., and Clementi, E.: Machine learning methods to predict Sea Surface Temperature and Marine Heatwave occurrence: a case study of the Mediterranean Sea, EGUsphere, 1-22,
665 10.5194/egusphere-2023-1847, 2023a.

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, 10.5194/essd-15-1269-2023, 2023b.

Chattopadhyay, A., Nabizadeh, E., and Hassanzadeh, P.: Analog Forecasting of Extreme-Causing Weather Patterns Using
670 Deep Learning, 12, e2019MS001958, https://doi.org/10.1029/2019MS001958, 2020.

Chen, L., Han, B., Wang, X., Zhao, J., Yang, W., and Yang, Z.: Machine Learning Methods in Weather and Climate Applications: A Survey, 10.3390/app132112019, 2023.

Coppini, G., Clementi, E., Cossarini, G., Salon, S., Korres, G., Ravdas, M., Lecci, R., Pistoia, J., Goglio, A. C., Drudi, M., Grandi, A., Aydogdu, A., Escudier, R., Cipollone, A., Lyubartsev, V., Mariani, A., Cretì, S., Palermo, F., Scuro, M.,
675 Masina, S., Pinardi, N., Navarra, A., Delrosso, D., Teruzzi, A., Di Biagio, V., Bolzon, G., Feudale, L., Coidessa, G., Amadio, C., Brosich, A., Miró, A., Alvarez, E., Lazzari, P., Solidoro, C., Oikonomou, C., and Zacharioudaki, A.: The Mediterranean forecasting system. Part I: evolution and performance, EGUsphere, 2023, 1-50, 10.5194/egusphere-2022-1337, 2023.

Darmaraki, S., Somot, S., Sevault, F., and Nabat, P.: Past Variability of Mediterranean Sea Marine Heatwaves, Geophysical
680 Research Letters, 46, 9813-9823, https://doi.org/10.1029/2019GL082933, 2019a.

Darmaraki, S., Somot, S., Sevault, F., Nabat, P., Cabos Narvaez, W. D., Cavicchia, L., Djurdjevic, V., Li, L., Sannino, G., and Sein, D. V.: Future evolution of Marine Heatwaves in the Mediterranean Sea, Climate Dynamics, 53, 1371-1392, 10.1007/s00382-019-04661-z, 2019b.

Darmaraki, S., Denaxa, D., Theodorou, I., Livanou, E., Rigatou, D., Raitsos E, D., Stavrakidis-Zachou, O., Dimarchopoulou,
685 D., Bonino, G., McAdam, R., Organelli, E., Pitsouni, A., and Parasyris, A.: Marine Heatwaves in the Mediterranean Sea: A Literature Review, Mediterranean Marine Science, 25, 586-620, 10.12681/mms.38392, 2024.

Desai, S. and Strachan, A.: Parsimonious neural networks learn interpretable physical laws, Scientific Reports, 11, 12761, 10.1038/s41598-021-92278-w, 2021.

Doury, A., Somot, S., Gadat, S., Ribes, A., and Corre, L.: Regional climate model emulator based on deep learning: concept
690 and first evaluation of a novel hybrid downscaling approach, Climate Dynamics, 60, 1751-1779, 10.1007/s00382-022-06343-9, 2023.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27, 861-874, https://doi.org/10.1016/j.patrec.2005.10.010, 2006.

Fdez-Riverola, F., Corchado, J. M., and Torres, J. M.: An Automated Hybrid CBR System for Forecasting, Advances in
695 Case-Based Reasoning, Berlin, Heidelberg, 2002//, 519-533,

Garcia-Gorriz, E. and Garcia-Sanchez, J.: Prediction of sea surface temperatures in the western Mediterranean Sea by neural networks using satellite observations, 34, https://doi.org/10.1029/2007GL029888, 2007.

Garrabou, J., Gómez-Gras, D., Medrano, A., Cerrano, C., Ponti, M., Schlegel, R., Bensoussan, N., Turicchia, E., Sini, M., Gerovasileiou, V., Teixido, N., Mirasole, A., Tamburello, L., Cebrian, E., Rilov, G., Ledoux, J. B., Souissi, J. B.,
700 Khamassi, F., Ghanem, R., Benabdi, M., Grimes, S., Ocaña, O., Bazairi, H., Hereu, B., Linares, C., Kersting, D. K., la Rovira, G., Ortega, J., Casals, D., Pagès-Escolà, M., Margarit, N., Capdevila, P., Verdura, J., Ramos, A., Izquierdo, A., Barbera, C., Rubio-Portillo, E., Anton, I., López-Sendino, P., Díaz, D., Vázquez-Luis, M., Duarte, C., Marbà, N.,

Aspillaga, E., Espinosa, F., Grech, D., Guala, I., Azzurro, E., Farina, S., Cristina Gambi, M., Chimienti, G., Montefalcone, M., Azzola, A., Mantas, T. P., Fraschetti, S., Ceccherelli, G., Kipson, S., Bakran-Petricioli, T., Petricioli, D., Jimenez, C., Katsanevakis, S., Kizilkaya, I. T., Kizilkaya, Z., Sartoretto, S., Elodie, R., Ruitton, S., Comeau, S., Gattuso, J. P., and Harmelin, J. G.: Marine heatwaves drive recurrent mass mortalities in the Mediterranean Sea, Global change biology, 28, 5708-5725, 10.1111/gcb.16301, 2022.

Giamalaki, K., Beaulieu, C., and Prochaska, J. X.: Assessing Predictability of Marine Heatwaves With Random Forests, 49, e2022GL099069, https://doi.org/10.1029/2022GL099069, 2022.

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, 172816-172829, 10.1109/ACCESS.2019.2955957, 2019.

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C. J., Benthuysen, J. A., Burrows, M. T., Donat, M. G., Feng, M., Holbrook, N. J., Moore, P. J., Scannell, H. A., Sen Gupta, A., and Wernberg, T.: A hierarchical approach to defining marine heatwaves, Progress in Oceanography, 141, 227-238, https://doi.org/10.1016/j.pocean.2015.12.014, 2016.

Hornik, K.: Approximation capabilities of multilayer feedforward networks, Neural Networks, 4, 251-257, https://doi.org/10.1016/0893-6080(91)90009-T, 1991.

Ibtehaz, N. and Rahman, M. S.: MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Networks, 121, 74-87, https://doi.org/10.1016/j.neunet.2019.08.025, 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, 10.1038/s41586-022-04573-9, 2022.

Jacques-Dumas, V., Ragone, F., Borgnat, P., Abry, P., and Bouchet, F.: Deep Learning-Based Extreme Heatwave Forecast, 4, 10.3389/fclim.2022.789641, 2022.

Kingma, D. P. and Ba, J. J. a. p. a.: Adam: A method for stochastic optimization, 2014.

Konsta, K., Doxa, A., Katsanevakis, S., and Mazaris, A.: Projected intensification of subsurface marine heatwaves under climate change, 10.21203/rs.3.rs-3091828/v1, 2023.

Lacoue-Labarthe, T., Nunes, P. A. L. D., Ziveri, P., Cinar, M., Gazeau, F., Hall-Spencer, J. M., Hilmi, N., Moschella, P., Safa, A., Sauzade, D., and Turley, C.: Impacts of ocean acidification in a warming Mediterranean Sea: An overview, Regional Studies in Marine Science, 5, 1-11, https://doi.org/10.1016/j.rsma.2015.12.005, 2016.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science (New York, N.Y.), 382, 1416-1421, 10.1126/science.adi2336, 2023.

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P.: Focal Loss for Dense Object Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 42, 318-327, 10.1109/TPAMI.2018.2858826, 2020.

Liu, J., Zhang, T., Han, G., and Gou, Y.: TD-LSTM: Temporal Dependence-Based LSTM Networks for Marine Temperature Prediction, 10.3390/s18113797, 2018.

Liu, Y., Weisberg, R. H., Sorinas, L., Law, J. A., and Nickerson, A. K.: Rapid Intensification of Hurricane Ian in Relation to Anomalously Warm Subsurface Water on the Wide Continental Shelf, Geophysical Research Letters, 52, e2024GL113192, https://doi.org/10.1029/2024GL113192, 2025.

Maas, A. L., Hannun, A. Y., and Ng, A. Y.: Rectifier nonlinearities improve neural network acoustic models, Proc. icml, 3,

Mavropoulou, A.-M., Mantziafou, A., Jarosz, E., and Sofianos, S.: The influence of Black Sea Water inflow and its synoptic time-scale variability in the North Aegean Sea hydrodynamics, Ocean Dynamics, 66, 195-206, 10.1007/s10236-016-0923-5, 2016.

McAdam, R., Masina, S., and Gualdi, S.: Seasonal forecasting of subsurface marine heatwaves, Communications Earth & Environment, 4, 225, 10.1038/s43247-023-00892-5, 2023.

McMillin, L. M.: Estimation of sea surface temperatures from two infrared window measurements with different absorption, Journal of Geophysical Research (1896-1977), 80, 5113-5117, https://doi.org/10.1029/JC080i036p05113, 1975.

Menna, M., Gačić, M., Martellucci, R., Notarstefano, G., Fedele, G., Mauri, E., Gerin, R., and Poulain, P.-M.: Climatic, Decadal, and Interannual Variability in the Upper Layer of the Mediterranean Sea Using Remotely Sensed and In-Situ Data, 10.3390/rs14061322, 2022.

Nguyen, Q. D. and Thai, H.-T.: Crack segmentation of imbalanced data: The role of loss functions, Engineering Structures, 297, 116988, https://doi.org/10.1016/j.engstruct.2023.116988, 2023.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M. J., Heinrich, M. P., Misawa, K., Mori, K., McDonagh, S. G., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D. J. A.: Attention U-Net: Learning Where to Look for the Pancreas, abs/1804.03999, 2018.

Oliver, E. C. J., Donat, M. G., Burrows, M. T., Moore, P. J., Smale, D. A., Alexander, L. V., Benthuysen, J. A., Feng, M., Sen Gupta, A., Hobday, A. J., Holbrook, N. J., Perkins-Kirkpatrick, S. E., Scannell, H. A., Straub, S. C., and Wernberg, T.: Longer and more frequent marine heatwaves over the past century, Nature Communications, 9, 1324, 10.1038/s41467-018-03732-9, 2018.

Parasyris, A., Alexandrakis, G., Kozyrakis, G. V., Spanoudaki, K., and Kampanis, N. A.: Predicting Meteorological Variables on Local Level with SARIMA, LSTM and Hybrid Techniques, 10.3390/atmos13060878, 2022.

Petrelli, P.: XMHW: Xarray based code to identify Marine HeatWave events and their characteristics, 10.5281/zenodo.6270280, 2022.

Pisano, A., Ciani, D., Marullo, S., Santoleri, R., and Buongiorno Nardelli, B.: A new operational Mediterranean diurnal optimally interpolated sea surface temperature product within the Copernicus Marine Service, Earth Syst. Sci. Data, 14, 4111-4128, 10.5194/essd-14-4111-2022, 2022.

Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 234-241,

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200097, 10.1098/rsta.2020.0097, 2021.

Sevault, F.: Atlas of the 1980–2018 ERA-interim simulation with the coupled regional climate system model CNRM-RCSM6 (version v2), 2024.

Sharma, S., Sharma, S., and Athaiya, A.: Activation Functions in Neural Networks, International Journal of Engineering Applied Sciences and Technology, 04, 310-316, 10.33564/IJEAST.2020.v04i12.054, 2020.

Smith, K. E., Burrows, M. T., Hobday, A. J., Sen Gupta, A., Moore, P. J., Thomsen, M., Wernberg, T., and Smale, D. A.: Socioeconomic impacts of marine heatwaves: Global issues and opportunities, Science (New York, N.Y.), 374, eabj3593, 10.1126/science.abj3593, 2021.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15, 1929-1958, 2014.

Sun, W., Zhou, S., Yang, J., Gao, X., Ji, J., and Dong, C.: Artificial Intelligence Forecasting of Marine Heatwaves in the South China Sea Using a Combined U-Net and ConvLSTM System, 15, 4068, 2023.

Taylor, J. and Feng, M.: A deep learning model for forecasting global monthly mean sea surface temperature anomalies, 4, 10.3389/fclim.2022.932932, 2022.

Velaoras, D., Zervakis, V., and Theocharis, A.: The Physical Characteristics and Dynamics of the Aegean Water Masses, in: The Aegean Sea Environment: The Geodiversity of the Natural System, edited by: Anagnostou, C. L., Kostianoy, A. G., Mariolakos, I. D., Panayotidis, P., Soilemezidou, M., and Tsaltas, G., Springer International Publishing, Cham, 231-259, 10.1007/698_2020_730, 2024.

Waldman, R., Somot, S., Herrmann, M., Bosse, A., Caniaux, G., Estournel, C., Houpert, L., Prieur, L., Sevault, F., and Testor, P.: Modeling the intense 2012–2013 dense water formation event in the northwestern Mediterranean Sea: Evaluation with an ensemble simulation approach, Journal of Geophysical Research: Oceans, 122, 1297-1324, https://doi.org/10.1002/2016JC012437, 2017.

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, 111358, https://doi.org/10.1016/j.rse.2019.111358, 2019.

800   Xu, Z., Xiao, Z., Zhao, X., Ma, Z., Zhang, Q., Zeng, P., and Zhang, X.: Derivation of Landslide Rainfall Thresholds by Geostatistical Methods in Southwest China, 10.3390/su16104044, 2024.

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, 2, e220089, https://doi.org/10.1175/AIES-D-22-0089.1, 2023.

805