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

Response to Reviewers

We thank the reviewers for their thoughtful and constructive feedback. Below, we provide detailed responses to their final comments, highlighted in blue, while retaining the reviewers' original text in black for reference. The line numbers mentioned correspond to the revised manuscript (clean version, without track changes)

RC #2:

I found your study very interesting and relevant to the growing field of AI-driven forecasting for extreme ocean events.

We thank the reviewer for their kind comments and are glad that they deem this paper valuable to the field. We believe addressing their comments has made this paper more robust.

Below are some comments that I believe could strengthen your manuscript:

1. Comparison with ocean models: While you acknowledge ocean models in the introduction, there is no direct comparison between your AI-based approach and a traditional numerical ocean model for MHW forecasting. A quantitative comparison—whether in terms of forecast accuracy, computational efficiency, or ability to capture physical processes—would provide valuable context and help clarify how AI complements or improves upon traditional approaches.

   We appreciate this suggestion. In AI-based approaches, it is standard practice to compare the predictions of the AI model using a testing subset. This constitutes a quantitative evaluation of the AI-based solution relative to the model's performance in predicting MHWs without AI. For this reason, a portion of the dataset is excluded from training, ensuring the model has never encountered those values. We have explicitly described this process in lines (209-214), and its results are reported in Section 3. Specifically, Fig. 4 and Tables 1–2 show the TPR, TNR, and FAR metrics, which are computed using the test dataset. These metrics compare the predictions from the ML model predictions to the "true" MHW labels, as these were derived from the RCSM model, by applying the MHW detection algorithm to the hindcast SST for all test days. The accuracy of these results is further discussed on lines 521-531 that read:

   “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 from the 3-day to the 7-day forecast scenario, 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.”

2. Resolution of the output: You mention that the final spatial resolution is downsampled to a 128x272 grid, but it would be helpful to explicitly state the corresponding resolution in kilometers. Additionally, how does the downsampling affect the model’s ability to capture finer-scale MHW dynamics, particularly in regions with complex topography such as the Adriatic or Aegean Seas?

Thank you for your point. Due to computational constraints, we were unable to train the model at the original 6-8 km resolution. Thus, we performed a downsampling to an approximate resolution of 12-16 km, which is now explicitly mentioned in line 146. While downsampling itself may introduce uncertainties, these are better discussed in the added lines (538-547) that now read: “*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 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*”.

A key limitation of neural network methodologies is the inability to precisely identify the source of each inaccuracy. Fine-tuning these ML models relies on trial and error and comparisons with alternative approaches, as demonstrated in the Persistence benchmark section (section 2.5) and the sensitivity analysis (section 3.2) conducted in this study.

For example, uncertainties in our results could also arise from excluding wind, and perhaps other atmospheric variables, from our input dataset, from the imbalanced input dataset, as well as from accumulated model errors in regions with complex topography. We now discuss these, in added lines (545-557), that read:

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 weak dependence of the SST on wind speed across all the Mediterranean sub-basins they considered, by calculating the Mutual Information index prior to applying the ML method. 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).

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). ”

3. Uncertainty Quantification: Your study provides a robust evaluation of model performance, but there is limited discussion on uncertainty quantification. Given the stochastic nature of neural networks, have you assessed the sensitivity of your predictions to different training datasets, hyperparameter choices, or initial conditions? Methods such as ensemble modeling could provide

insights into the confidence of the forecasts.

We appreciate the reviewer's comments regarding uncertainty quantification. As highlighted in Tables 1 and 2, our study performs a sensitivity analysis where models were retrained with different numbers of input variables. In most cases, the stability of the TPR, TNR, CM and FAR metrics remains high, except when an essential variable is entirely removed. We believe this provides a measure of robustness regarding input data choices. Additionally, in the supplementary material we show results for a different training-testing dataset splitting selection and we comment on the results on lines 559-573 that now read as:

"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 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. "

Regarding different initial conditions, we do not have control over the circulation models which were used as inputs for our ML case study. These were kindly provided by the Med-CORDEX initiative, described in section 2.2, and we could not alter initial conditions to retrain with different models.

While ensemble modeling would, indeed, offer further insights into prediction confidence, computational constraints limited our ability to explore this in the current study. We now acknowledge and discuss this as a potential avenue for future research in lines 612-620, that read as:

"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."

4. Generalization Beyond the Mediterranean: You mention that the proposed method could be applied to other regions, but it would be helpful to discuss potential challenges in doing so. For

> example, would a model trained on Mediterranean SST anomalies generalize well to other basins with different oceanographic characteristics (e.g., stronger currents, different stratification, or more extreme variability)? A brief discussion of how the model could be adapted or retrained for different environments would be valuable.

We thank the reviewer for raising this point. The current trained CNN model is specific to the Mediterranean region due to the spatiotemporal characteristics of the dataset used for training. However, the methodology itself is adaptable and could be retrained using region-specific datasets. Given the Mediterranean Sea is a region where diverse circulation and thermohaline patterns can be observed, we can infer that the methodology can be effectively applied elsewhere. For instance, a similar AI technique has already been employed in the South China Sea (Sun et al., 2023), though without incorporating the novel loss function and attention mechanisms introduced in our work. We have added and amended a part of the discussion on this aspect to acknowledge the challenges and possibilities of generalizing our approach to other regions. (lines 586-593) that now read as:

“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. “

Overall, this study presents a promising application of deep learning for MHW forecasting, and I appreciate the detailed methodology and validation process. Addressing these points could further enhance the robustness and impact of your work.

We thank the reviewer for their positive feedback. We believe that addressing their feedback has significantly strengthened the robustness of the paper.

References:

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, 2023, 1-22, 10.5194/egusphere-2023-1847, 2023.

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.