Review Ocean Science *Bonino et al 2023 :*
Machine learning methods to predict Sea Surface Temperature and Marine Heatwave occurrence: a case study of the Mediterranean Sea

General comments

In this manuscript, the authors aim to develop a short-term SST and MHW framework based on contemporary machine learning (ML) techniques, using the Mediterranean Sea as an illustrative example. This subject addresses important and urgent issues concerning the increasing number of marine heatwaves affecting this area. To achieve this goal, the authors evaluate three common ML techniques: Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Random Forest (Rforest).

This topic propose a method to bridge a gap in the operational prediction of MHW using a low computational data-driven approach.

In my opinion, this work is interesting and demonstrates the adaptability of these techniques to such an urgent topic. The manuscript is well-written and easy to read. However, after reading the manuscript, I find myself with more questions than answers about this study, mainly due to a lack of explanations regarding the construction of the framework and the data used. In particular, these concerns are about two topics:

First, the construction of the ML classification for extreme events. By definition, an extreme event is a phenomenon that is rare in a dataset, making it challenging to identify patterns and leading to imbalanced classes. In my opinion, this situation may explains the high number of FPR you found in the MHW prediction. This question about extreme events is crucial in the context of climate variations driven by climate change. The data used in this study seem not to be detrended, allowing extremes to originate from different sources, even when only meteorological predictors are considered. It would be beneficial to have a smaller sample in the training dataset with an acceptable bias to provide the network with more information about extremes (it a common way of doing, for example you can refer to the work by Mounier et al. 2022).

Mounier, A., Raynaud, L., Rottner, L., Plu, M., Arbogast, P., Kreitz, M., ... & Touzé, B. (2022). Detection of bow echoes in kilometer-scale forecasts using a convolutional neural network. *Artificial Intelligence for the Earth Systems*, *1*(2), e210010

The second concern relates to the study's objectives. If I read correctly, the justification of this study is the need for a prediction system for SST and MHW due to their significant social, ecological, and economic impacts: I couldn't agree more. However, I did not find any justification, and I'm not convinced by, the impact of having a short-term forecast of SST/MHW. Given the profound impact of MHWs on marine ecosystems and, consequently fisheries, and the timescales of MHWs (lasting for days to years) many researchers are focusing more on multi-weeks or seasonal forecasts to anticipate these impacts. I can speculate on why you developed this short-term solution, but it is not clearly justified in the article and seems to be useful for very specific industry. For example, does an extreme 1week MHW has more impact than a moderate MHWs that lasts 1 year. In connection with this, I would suggest exploring a comparison between these ML techniques and multiple linear regression using the same predictors. This would help justify choosing ML over a statistical model with lower computational costs, as you mentioned in Page 2, Line 42.

All of these points highlight the need for a more comprehensive discussion and justification of this study, which undoubtedly has some very interesting findings. Despite the length of this review, I'm convinced by the importance of such work and the publication in Ocean Science.

Line by line comments

Abstract

L13:
Regarding your results, you could not conclusively state that ML techniques 'outperform' MedFS, even in the context of 3-day predictions. In my opinion, the results fall within the margin of error (please refer to my comments in the Results section). Furthermore, you should replace 'most regions' with a quantitative assessment of this outcome.

I also suggest adding a closing sentence to your abstract to provide some perspective. However, the decision is up to you.

Introduction

As mentioned in the general comments, the introduction lacks justification for considering only a 7-day forecast period. If we compare this to other scientific fields, such as numerical weather prediction, the justification for using ML techniques often leans in the opposite direction: attempting to outperform models beyond the 7-day threshold, where predictability reaches its limits. In my opinion, this approach would be even more compelling for SSTs, given the relatively small range of variation within a week, especially in your case when utilizing interpolated L4 data for training.

P1.L18-21

I would recommend rewriting the introduction to separate the sections addressing SST, the impact on ecosystems, and the definition of MHWs as extreme events. In my opinion, there is a bit of confusion (at least as a reader) between these concepts in the initial part of the introduction, and you are defining them a second times in the Results section (Page 8, Line 229).

P2.L39
I'm not familiar with the reference you mentioned but I have in mind one counter example. In numerical weather prediction AI models like FourCastNest, the model provides forecasts globally for up to 15 days using ML techniques exclusively. This raises an important questions about the applicability of your study, especially given the wide area you are examining. Following this sentence it would be preferable to use numerical approach in the case of Med Sea and I could not find the justification in your introduction.

P2.L44-65

I think you could shorten this section which is currently a very long review. As you mentioned, statistical techniques have a well established history in SST including bias estimation and satellite data reconstruction.

Saux Picart, S.; Tandeo, P.; Autret, E.; Gausset, B. Exploring Machine Learning to Correct Satellite-Derived Sea Surface Temperatures. Remote Sens. 2018, 10, 224. https://doi.org/10.3390/rs10020224

Barth, A., Alvera-Azcárate, A., Licer, M., and Beckers, J.-M.: DINCAE 1.0: a convolutional neural network with error estimates to reconstruct sea surface temperature satellite observations, Geosci. Model Dev., 13, 1609–1622, https://doi.org/10.5194/gmd-13-1609-2020, 2020

Methodological Framework

- One of my main concerns in this section deals with the choice of the data and its inherent justification. If I understand correctly, the aim of the study is to develop a SST forecast up to 7 days. However, L4 ESA CCI SST product is primarly designed for climate studies and is a gap-free product, meaning that some of the data are computed using interpolation. By using such data, you may inadvertently introduce biases into your model. For example, data are smoothed and you may hide upwelling in coastal areas.
  An alternative approach could have been to use dedicated L3 operational data (near real time product), such as those developed by the same ESA CCI project or the OSI SAF project.

- In connection with this, how do you account for the differences in spatial resolution between the ESA CCI and ERA5 datasets? You are attempting to link SST variations to atmospheric predictors but ERA5 has a spatial resolution of 0.25° whereas ESA CCI provides SST products at a resolution of 0.05°.
  I might have missed something in the study but it appears to me that you are calculating a mean SST for each area. Do you have a single SST representative for each zone?
  I have the same question with the atmospheric variables. However these zones represent large areas and by aggregating data in this way you may smooth out the data. Thereby by neglecting the non-linearity of wind speed, which is a key driver of SST variations especially in the Mediterranean Sea where you have regional winds that interact strongly with SSTs.

- I'm also surprised that you did not account for 2m air surface temperature, specific humidity and mixed layer depth variables as they are known drivers of MHWs and even more in the Mediterranean Sea. You took T2m and q2m into account in the sensible/latent heat fluxes but it would be interesting to know about their direct impact. I'm also referring to the MLD because MHWs often have a vertical extension and can persist in intermediate layers without necessarily having any signature at the surface. See:

Amaya, D. J., Miller, A. J., Xie, S.-P., and Kosaka, Y.: Physical drivers of the summer 2019 North Pacific marine heatwave, Nat. Commun., 11, 1903, https://doi.org/10.1038/s41467-020-15820-w, 2020.

Chen, HH., Wang, Y., Xiu, P. *et al.* Combined oceanic and atmospheric forcing of the 2013/14 marine heatwave in the northeast Pacific. *npj Clim Atmos Sci* **6**, 3 (2023). https://doi.org/10.1038/s41612-023-00327-0

Guinaldo, T., Voldoire, A., Waldman, R., Saux Picart, S., and Roquet, H.: Response of the sea surface temperature to heatwaves during the France 2022 meteorological summer, Ocean Sci., 19, 629–647, https://doi.org/10.5194/os-19-629-2023, 2023

Eric C.J. Oliver, Jessica A. Benthuysen, Sofia Darmaraki, Markus G. Donat, Alistair J. Hobday, Neil J. Holbrook, Robert W. Schlegel, Alex Sen Gupta, Annual Review of Marine Science 2021 13:1, 313-342

- One of the advantages of the numerical approach is its ability to account for a wide variety of processes that are parameterized within the model. However, here you are only exclusively studying SST predictions within the scope of the atmospheric forcings. How do you account for oceanic processes? Given their potential significance? For example, smoothing wind speed variations might reduce the contribution of the vertical mixing among other crucial oceanic processes. You mention about this in your discussion but this could be extended.

- Regarding the detection of MHWs detection, the methodology is not entirely clear for me. Firstly, I could not find any explanation of how you calculate the daily climatology and the seasonal threshold. Are these values unique for each regions or are they calculated for every pixels at the ESA CCI resolution? Do you consider an MHW to occur when at least 1 pixels within a region exceeds the threshold or do you consider the mean SST over the region?
  Furthermore, calculating a climatology by only taking the mean of daily SSTs over 30 years can introduce some biases because it may not be statistically robust and sensitive to extremes. Typically the climatology is computed for a specific day using 30 years of data either with an 11-day sliding window centered on the day in question or by employing the first harmonic of a Fourier series.
  Do you account for situations when a MHWs occurs less than 2 days before the end of another one it is considered part of the MHW?
  P6.L175: you mention computing the climatology over the period 1981-2016 which is not the international standard of using 30 years. It is worthy noting that the first complete year of ESA CCI data is 1982. More generally, you should clarify how you calculate the climatology and anomalies.

- In studying MHWs, it is important not only to detect them but also to estimate key metrics such as the mean intensity, max intensity and the severity. Do you have insights into the possibility of predicting these metrics using your model?

P4.L104
This part should be moved to the results and discussions sections. Moreover, GEO and INC are not independent variables, which may explains why they are correlated with SST. Even though you are looking at daily mean, ESA CCI SST are representative of the foundation temperature which is not influenced by the diurnal cycle. Therefore, you should rephrase the sentence ‘INC influenced directly SST during daily time’ (P4.L7). I understand what you meant and suggest this modification to avoid the reader to misunderstood.

I’m not entirely convinced that MI calculations are necessary to identify relations between variables. Examining the upper-ocean layer equation could give the same result (although it is not quantitative). Additionally, it is excepted to find a weak correlation between wind speed and SST due to the non-linearity I mentioned earlier which you did not take into account.

P5.L124-141

I have several questions regarding the description of the networks you used.

Firstly, you have not detailed the inputs/ouputs and how the variables are incorporated in the networks (possibly as concatenated large vectors?).
Additionally, I found no indications in this work regarding how the data were normalized to make different atmospheric and oceanic variables comparable.
Additionally, I find it somewhat surprising that you used 200 epochs for both LSTM and CNN without discussing the complexity of the networks. For instance, in atmospheric models, LSTM converge quickly compared to CNN and it is common practice to add an early stopping to prevent overfitting.

P6.L158
This is a minor comment but you mentioned computational time without providing hardware specifications. I may find this information useful.

Equation 4: There is a typo in the F1 score, it is 2*P*R/(P+R)

Results

- I'm a bit surprised by the lack of stability in the ML techniques (this might be related to a lack of experience with these techniques from my side). On the contrary, as you mentioned in P7.L210, it is normal for the dynamical model to be rather stable for up to 7 days (in my opinion it may depends on the predictability limits). Are you sure that the behavior of your ML models is not influenced by some sort of persistence or memory effect that could explain the increase in RMSE? I thought about this because you may have introduced some biases during the training stage, as I mentioned earlier. Additionally, you are also studying SST through a spatial average over large areas. Furthermore, at the first order, a 0.1°C difference is within the common measurement errors in satellite SST products.
  Regarding all this, I would not claim that ML techniques 'outperformed' the numerical model, but as you mentioned in P8.L213, they 'compare favorably'.

- In a second step, your study and the results are quite interesting. In my opinion, it could be improved by an in-depth analysis of the results by regions, attempting to understand the diversity of responses from different basins (explaining it through dynamic conditions or other factors). For example, the dynamic in regions like the Alboran Sea may explain most of the variability. I understand that it may not be feasible to include everything in a single study, but this could be explored in the context of an additional study.

- **Figure 5** is very informative, would it be possible for you to add the daily climatological mean in addition to the threshold. It is just a suggestion but it would be also interesting to add a focus on a particular MHWs because it seems that FigS3 shows some sort of time lag in the SST between models. Maybe add some metrics such as the correlation, annual mean and standard deviation.

- **Figure 6:** I found it a bit difficult to read the figure with two distinct labels for SST and the other variables. In addition, the lines are very thin, and I can't distinguish easily between the variables.

Discussion/Conclusion

In general, the discussion is good, with some very interesting points and references. I appreciate that you are not attempting to oversell your results. ML techniques, despite some limitations, have demonstrated predictive skill, and your discussion about the challenges in understanding MHW, along with a thoughtful exploration of time-scales, raised some intriguing questions.

L293-295: In regard of your results the ML techniques seems to have better result however regarding the stability, I'm not convinced that it is not linked to some memory effect. I'm also not sure of the impact of predicting SSTs at 1 and 3 days, except for very specific industry with high tolerance to the risks.

L310: I agree with the contribution of incoming solar radiation, which is generally linked to lower than average anomalies in cloud cover. However, this contribution is limited in summer in the Mediterranean Sea due to the usual low cloud cover over the area. Thus, MHWs are primarily driven by other variables, such as heat fluxes (namely the atmospheric variables T2M, Q2M and WS) which exhibit significant regional dependencies (Guinaldo et al., 2023, as mentioned earlier). Additionally, you could have also discussed the possibility of incorporating ocean heat content to improve the forecast (Holbrook et al., 2020).

Another important point to consider is related to the data you used. As mentioned earlier, you employed SST data dedicated to climate studies instead of near-real-time data. Your discussion would benefit from addressing the limitations associated with using such data. In the context of the study framework, it would also be worthy to discuss the comparison between this type of forecasting model and other approaches, such as multiple linear regression or model ensembles.

To enhance this discussion, I recommend reading the following study:

Benthuysen, J. A., Smith, G. A., Spillman, C. M., & Steinberg, C. R. (2021). Subseasonal prediction of the 2020 Great Barrier Reef and Coral Sea marine heatwave. *Environmental Research Letters*, *16*(12), 124050.