Egusphere-2023-1847

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

Authors: Bonino et al., 2023

Scientific significance:
The manuscript describes a set of experiments obtained by training shallow (Random Forest, RForest) and deep learning (Long short-term memory, LSTM and Convolutional Neural Network,CNN) methods using ESA CCI SST and ECMWF datasets. The main goal is to predict Sea Surface Temperature and Marine Heat Waves in the Mediterranean Sea. The duration of the prediction is 7 days. Results are provided by using daily RMSE values.

Scientific quality:
The scientific question raised in this study is clear (Line 6 "*create competitive predictive tools"* and Line 69 *provide a proof-of-concept study on the advantage of data-driven ML methods*).
The work is relevant for the novelty of the aim, of the methods and for its reversibility in being used in different scientific fields.

With the advancement of NN algorithms, it is important to verify whether and to what extent data-driven methods can provide comparable products with respect to classical numerical methods.
Although the scientific contribution of this study is significant, some scientific choices need further explanations. In my opinion, the manuscript (especially the introduction) lacks useful information. This could diminish the importance of the work, especially of the methods, which I found particularly robust.

Presentation quality:
The manuscript is generally well presented and well written. Since the scientific value of this study is high, my suggestion is to add information that will enrich your study and efforts. In the *General comments* below, I have tried to list them. I hope it will be useful.

General comments:
I believe that some scientific choices need further explanations.

1) I suggest answering the following questions by adding the information in the INTRODUCTION

Both NNs and numerical models are able to cover 7 days of forecast. While the NNs take less computational time, the numerical models have higher accuracy in the last days of the forecast (SST).

- What are the main advantages of using NN to predict ocean physics? Are the advantages only computational ones? (NN can approximate nonlinear function (Hornik et al., 1989)

- What are the main limitations in applying NN to real-world scenarios? (e.g. NN algorithms do not know when they are violating the laws of physics, e.g. Buizza et al., 2022)

The goal of the work is to "*create competitive predictive tools*". You use the RMSE to assess the feasibility of NNs in predicting SST and MHWs and you validate all the NN tests (also) comparing the results with a numerical model.

- In this case I believe that the Figures (time-series) you have chosen are sometimes contradictory and sometimes do not show what you are describing (see detailed comments)

Some interesting information on this work can be found in the discussion. I expected to read them in the introduction (see detailed comments).

2) I suggest answering the following questions by adding the information directly in the METHODS section:

In the text I have the impression that you are sometimes contradicting your scientific choices. For example when you write: "*For instance, the numerical approach is better suited for predictions over a wider area, while the data-driven techniques are more applicable for location-specific studies".* Based on this:

- Why are you proposing predictions *over a wider area with data-driven techniques*?

Since the MedFS forecast used for validation covers 10 days:

- Please explain why you chose a time interval of 7 days for the forecast (instead of 10 days).
- Have you considered or tried to extend the duration of the prediction?
  Following your results, I can infer that by extending the duration of the experiment (10days) for SST:
  NN's RMSE >> MedFS RMSE
  NN's computational time << MedFS computational time
  (please specify the computational time of the model in the text).

3) I suggest answering the following questions by adding the information directly in the DISCUSSION section:

I have the impression that the reader should follow the NN "branch" instead of the numerical models (to predict SST and MHWs). But "NN algorithms do not know when they are violating the laws of physics" . On the basis of this:

- Do you think that this limits the NNs-skills along the prediction interval?
- Have you considered adding physical constraints to your NN models?
- Have you considered the possibility of predicting SST and MHW by modifying the inputs of the NNs? (e.g. by adding ocean variables to atmospheric variables)

- Would you suggest using NNs to predict SST and MHWs in a 3/5-days interval and MedFS for a longer period?

Buizza, Caterina, et al. "Data learning: Integrating data assimilation and machine learning." *Journal of Computational Science* 58 (2022): 101525.

Detailed comments (minor):

Abstract:

To facilitate the reader, you might explicitly write the number of the Experiments.

- L 4: MHWs acronym already introduced in Line 1
- L 9: Following the outline of the manuscript I think it could be clearer to insert in L9 the sentence L 15-16
- L 11 typo error Cat °C at
- L 11 and CNN RMSE?

Introduction

I appreciated the description of the impact of MHWs on ecosystems, but since the article does not deal with ecological studies I would expect to have: (i) less information on the ecological impacts (1-2 lines) (ii) more information on the techniques chosen and their pros and cons, including a comparison with numerical models (iii) a brief overview of the main characteristics of the Mediterranean Sea with respect to the objective of your work (e.g The reason why you have chosen the Mediterranean Sea as your study area?). Please describe if there is a sector (western-central-eastern) that is more susceptible to these events.

About 25% of your abstract focuses on NN's improvements compared to numerical models (MedFS) which you mention in P9. I suggest adding a sentence on the MedFS model used to compare the results (at line 70).

- L 18-29: I think this part could be shortened.
- L 23: Garrabou et al., 2022 double citation or missing bibliography ?
- L 33: typos error *generations(Leroux*
- L 39: Explain your motivation with respect to your work: "..*data-driven techniques are more applicable for location-specific studies.".*
- L 40-41: I suggest removing the sentence because it seems to be out of topic
- L 55 typos error *begging*
- L 44-65: If in your opinion it is important, please highlight if there is a method (shallow/deep) that better fits the SST/MHWs prediction goals (from literature).
- L 44-65: Too many citations. I suggest reducing the number of citations and adding more detailed information (i.e. the duration of the predictions from the cited literature). Moreover, in many works you cited (Corchado ,Liu, Xie etc.) the duration of the prediction is higher than 7 days.

Methodological Framework:
Data collection and preprocessing

- L 83: Add the spatial resolution of the datasets e.g. “daily satellite-derived Sea Surface Temperature (SST) data”.
- L 86: the L4 dataset provides interpolated data (L3 does not). I assume that you have decided to use the L4 dataset to numerically enrich the dataset for training and test evaluation. Is this the case?
- L 102-114: In my opinion, table 1 is a result (as you write in the abstract L15-16)

Experiments/Evaluation metrics :

- L 147: I would have introduced MedFS earlier in the text (e.g. in the Introduction)
- L 173: remove citation in parentheses “*in Hobday et al. 2016 (Hobday et al., 2016)*”
- L. 182: you refer to a table where I guess there’s a typo in the first column-second row cell: *MH predicted*

Results:
If you are going to redo some figures, please increase the font size (especially for the legend).

- L 198: “For instance, in all the techniques, region 15 shows the lowest, or almost the lowest (e.g in CNN) RMSE, while region 11 shows the highest errors”
  I cannot recognise region 15 in the CNN subplot, it is hardly readable. From figure 3 I can only evaluate the daily variability of RMSE across the Med. basins. Perhaps consider that plotting the average RMSE would be more useful to have a general overview on which method 'outperforms' the others.
- L. 198-200: Is there a reason why RMSE in region 11 is higher and 15 lower? A brief overview of the main physical characteristics of the Mediterranean Sea in the introduction might be helpful.
- L. 205: “*They also represent different dynamical areas of the Mediterranean basin.”*
  Which ones? Add information in the Introduction.
- L. 210: “*In contrast to ML methods, the dynamical model’s prediction of SST is influenced by atmospheric forecasts throughout the forecast period, which likely prevents the RMSE from increasing with the lead time*.”
  I think this is not a result, I would rather move it into the Discussion section.
- L 215: “*ML methods show lower RMSE than the MedFS forecast system for the first 3 days of forecast and they are comparable at lead time of 5 days.”*
  From Figure 4a
  - it results that the RMSE of MedFS at day3 is less than the one from NN methods in west and east only.
  - why are MedFS bullets plotted only for days 1,3,5?
  - UEXPs are called whiskers in the Figure and bar / UEXPs in the text. Better to be consistent with the nomenclature.
- L 221: *“It is likely connected to the fact that CNN algorithms are typically designed for image processing rather than time-series forecasting”.* It would probably be better to have this information right from the introduction.

- L 223: Is there a particular reason to choose the year 2020?
- L 225: *The figure shows a very close match between the forecasts and the observations.* Hence NNs don't outperform the MedFS.
- L 232: "*while Figure 4b shows the variation of the F1 score for all the methods with increasing forecast lead time"*
  The reader expects at this point an explanation of the figure 4b. It comes at L244. I suggest removing this sentence or adding information previously.
- Figure 6: impossible to see the lines. Since you only explain INC SST and LAT why plotting the others atmo forcing? (you can merge all the other ones in 1 single line)
- L264 and L267: add abbreviation in the text (INC in fig… lat in..)

Discussion and conclusions:

Can you introduce possible future developments?

- L278-282: All the abbreviations are already introduced
- L299-301: These features are typical of all the NNs methods. I would prefer to have this general info in the Introduction section.
- L330-336: Too much information on the impact, I suggest reducing this paragraph.

Supplementary:

Correct the description of Figure S4
Figure S4: As Figure S4 but for forecast lead time 3.