Review of “Baltic Sea Surface Temperature Analysis 2022: A Study of Marine Heatwaves and Overall High Seasonal Temperatures” by Lindenthal et al. for the 8th edition of the Copernicus Marine Service Ocean State Report (OSR 8)

- **SUMMARY**

The present work conducts an analysis of the Marine Heatwaves (MHWs) detected in the Baltic Sea during the year 2022. To achieve this, the authors utilize various observational and modeling databases. The paper discusses the obtained results, focusing on the analysis of parameters characterizing the variability of MHWs throughout the year (intensity, frequency, and duration). The study describes the magnitude of MHWs in the Baltic environment and draws other interesting conclusions, such as the emergence of positive trends and the relationship between the vertical propagation of MHWs and the development of cold intermediate layers.
Hence, the preprint holds scientific value and falls within the scope of the Ocean State Report. However, some aspects requiring improvement have been identified for manuscript publication. These changes do not alter the substance of the work, allowing its publication with the implementation of minor revisions. Below is a list of the main reasons supporting this recommendation.
As a scientific reviewer, I am compelled to ensure the scientific quality of the contribution. Thus, despite being aware of the space limitations imposed on OSR submissions, some of the recommendations provided may conflict with these constraints. I leave it to the editor's discretion to decide on the implementation of such changes.

We thank the reviewer for their thoughtful and thorough review and helpful suggestions. We addressed the reviewer comments below in blue.

- **GENERAL COMMENTS**

1) The paper employs a considerable amount of geographical terminology (Bothnian Sea, Bothnian Bay, Baltic Proper, Gulf of Finland and Gulf of Riga).The use of this terminology is enriching and aids in the writing process; however, it should be noted that potential readers may not be familiar with Baltic geography. Therefore, I believe that the inclusion of an initial figure (map) displaying relevant data from the study area would be highly beneficial. This map could encompass the following information: Delimitation of the geographical zones used in the article, the location of observational stations (the markers in Figure 2 are hardly visible), bathymetry, etc.

   We will provide a new Figure with map and station and place names.

2) In the Introduction, I miss a clear statement of the objectives and motivation of this work.

   We will update the introduction and also the discussion in order to clarify our motivation and objectives in the revised manuscript.

3) One of the weakest points of the paper is related to the model validation (Section 2.4). While the validation exercise is appropriate,it is done inadequately in the paper.

   For the revised version of the section we will provide a more comprehensive model validation using k-means clustering algorithm (Raudsepp and Maljutenko, 2022), but move the details of this to the supplementary material in agreement with the editor. This method has been used for validation of the physical and biogeochemical models by Kõuts et al. (2022), Raudsepp et al. (2022, 2023).

Kõuts, M., Maljutenko, I., Elken, J., Liu, Y., Hansson, M., Viktorsson, L., Raudsepp, U., 2021. Recent regime of persistent hypoxia in the baltic sea. Environmental Research Communications, 3 (7), 075004. doi: 10.1088/2515-7620/ac0cc4

Raudsepp U, Maljutenko I. 2022. A method for assessment of the general circulation model quality using K-means clustering algorithm: a case study with GETM v2.5. Geosci Model Dev. 15:535–551. doi:10.5194/gmd-15-535-2022.

Raudsepp, U., Maljutenko, I., Haapala, J., Männik, A., Verjovkina, S., Uiboupin, R., von Schuckmann, K., Mayer, M., 2022. Record high heat content and lowice extent in the Baltic Sea during winter2019/20. In: Copernicus Ocean State Report, Issue 6, Journal of Operational Oceanography,15:sup1, s175–s185; DOI:10.1080/1755876X.2022.2095169

Raudsepp, U., Maljutenko, I., Barzandeh, A., Uiboupin, R., and Lagemaa, P., 2023. Baltic Sea freshwater content, in: 7th edition of the Copernicus Ocean State Report (OSR7), edited by: von Schuckmann, K., Moreira, L., Le Traon, P.-Y., Grégoire, M., Marcos, M., Staneva, J., Brasseur, P., Garric, G., Lionello, P., Karstensen, J., and Neukermans, G., Copernicus Publications, State Planet, 1-osr7, 7, https://doi.org/10.5194/sp-1-osr7-7-2023.

Supplement

We utilize a clustering method to assess the accuracy of the hydrodynamic model. This method provides insights into the overall model accuracy by clustering the errors. The clustering process employs the K-means algorithm, which is a form of unsupervised machine learning (Jain, 2010). The original description of this method can be found in the work of Raudsepp and Maljutenko (2022). In our assessment, all available data within the model domain and simulation period are included, even if the verification data is unevenly distributed or occasionally sparse. This approach allows us to evaluate the model quality at each specific location and time instance where measurements have been obtained.

The first step of the method is the formation of a two-dimensional error space of two simultaneously measured parameters. A two-dimensional error space (dS,dT) of simultaneously measured temperature and salinity values was formed as the basis for the clustering, where dS=(S_{mod}-S_{obs}) and dT=(T_{mod}-T_{obs}) with the model and observed salinity, S_{mod} and S_{obs}, and the model and observed temperature, T_{mod} and T_{obs}. The dataset utilized in this validation study was obtained from the EMODNET dataset compiled by SMHI (SMHI, 2019). It comprises a total of 3,094,089 observations that align with the simulation period of the Baltic Sea physics reanalysis (product ref. no. 3 in Table 1), encompassing the years 1993 to 2022. We extracted the nearest model values from the reanalysis dataset for each observation.

The second step is the selection of the number of clusters. For simplicity, we preselected five clusters. The third step is to perform a K-means clustering of the two-dimensional errors. The clustering is applied to the normalized errors. Normalization was done for temperature and salinity errors separately using corresponding standard deviations of the errors. The K-means algorithm finds the location of the centroids of a predefined number of clusters in the error space. The location of the centroids represents the bias of the set of errors for each cluster. The fourth step is calculation of statistical metrics of non-normalized clustered errors. Common statistics like standard deviation (STD), root mean square deviation (RMSD) and correlation coefficient can be calculated for the parameters belonging to each cluster.

The fifth step is the analysis of spatio-temporal distributions of the errors belonging to different clusters. In the formation of the error space, we retained the coordinates of each error point (*dS,dT*)(*x,y*), which enables us to map the errors belonging to each cluster back to the location where the measurements were performed. In order to do that, the model domain is divided into horizontal grid cells (*i,j*) of 27x27 km^2 size. Then the number of error points belonging to different clusters at each grid cell (*i,j*) is counted. Total number of error points belonging to the grid cell (*i,j*) is the sum of the points of each cluster. The share of error points in each grid cell belonging to cluster *k* is the ratio of the number of error points of cluster *k* and the total number of error points in each grid cell.

Figure A1 displays the results of the K-means clustering for non-normalized errors. Table S1 presents the corresponding metrics. Within cluster *k*=5, the salinity and temperature values closely align with the observations, with a bias of *dS*=-0.40 g/kg and *dT*=-0.02 °C, respectively. This cluster encompasses 57% of all data points. The points are distributed throughout the entire Baltic Sea, with a dominant share exceeding 0.5 (Figure S1b). Clusters *k*=3 and *k*=4 exhibit relatively even spatial distributions over the Baltic Sea, accounting for 11% and 8% of the points, respectively. These clusters are particularly noteworthy due to their relatively high temperature biases and variability, which are crucial for calculation of marine heatwaves. The clusters *k*=1 and *k*=2 represent the points with low temperature but high salinity error (Table S1). Spatially these points have high share in the southwestern Baltic Sea (Figure S1b) pointing to the occasional underestimation or overestimation of the inflow/outflow salinity.

Collectively, approximately 82% of all validation points exhibit relatively low temperature bias, STD, and RMSD (Table S1). The surface layer validation shows that less than 10% of comparison points have significant temperature errors (Figure S1c). Due to a low share of these validation points we do not expect the significant impact to the determination of the surface MHW and their statistics. Below the surface layer i.e., from a depth range of 0.5-40 m, the share of the clusters *k*=3 and *k*=4 could increase to 25% (Figure S1c). Consequently, we anticipate that the model reanalysis data provide sufficiently accurate information for calculating the subsurface MHW and their statistics of the Baltic Sea.

References

Jain, A. K. 2010. Data clustering: 50 years beyond K-means, Pattern Recognition Letters, 31, 651–666, doi:10.1016/j.patrec.2009.09.011

Raudsepp U, Maljutenko I. 2022. A method for assessment of the general circulation model quality using K-means clustering algorithm: a case study with GETM v2.5. Geosci Model Dev. 15:535–551. doi:10.5194/gmd-15-535-2022.

SMHI, 2019. Baltic Sea – Eutrophication and Acidity aggregated datasets 1902/2017 v2018, Aggregated datasets were generated in the framework of EMODnet Chemistry III, under the support of DG MARE Call for Tender EASME/EMFF/2016/006 – lot4, EMODnet Chemistry [data set], doi:10.6092/595D233C-3F8C-4497-8BD2-52725CEFF96B.

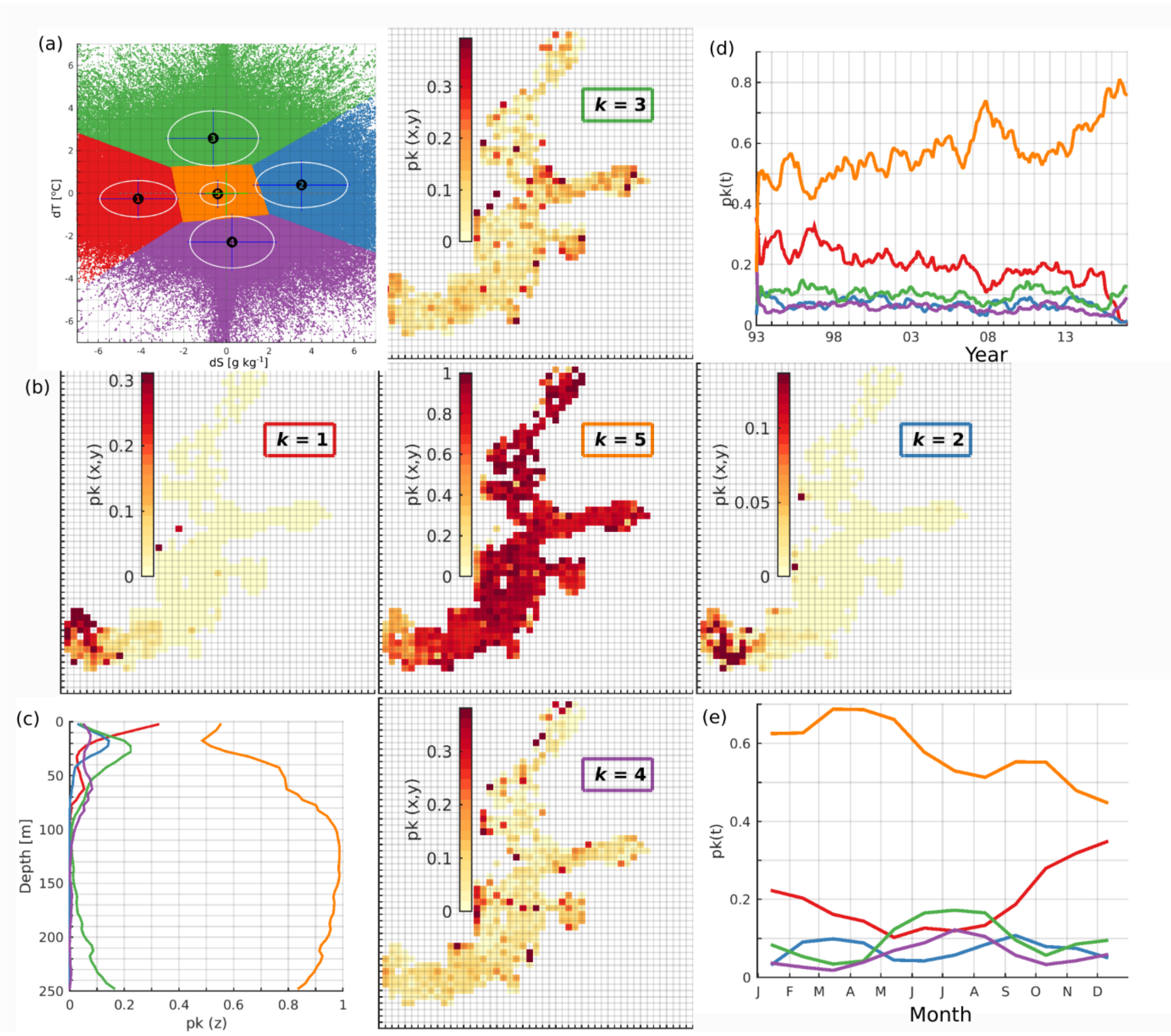

Figure S1. Distribution of normalized error clusters for K=5 (a). The spatial distribution (b, shaded sub-plots), vertical distribution (c), temporal distribution (d), and seasonal distribution (e) of the share of error points belonging to the five different clusters.

Table S1. The share (%), bias, root mean square deviation (RMSD), standard deviation (STD) and correlation coefficient (CORR) for each of the five clusters.

		BIAS		STD		RMSD		CORR		
k	share	dS	dT	dS	dT	S	T	S	T	dSdT
1	18	-4.137	-0.259	1.797	0.851	4.511	0.890	0.899	0.782	-0.094
2	7	3.531	0.389	2.160	1.061	4.140	1.130	0.930	0.750	-0.109
3	11	-0.618	2.581	2.123	1.284	2.211	2.882	0.965	0.584	-0.059
4	8	0.269	-2.291	1.972	1.214	1.990	2.593	0.954	0.709	-0.138
5	57	-0.401	-0.020	0.832	0.535	0.924	0.536	0.993	0.887	0.074

4) Additionally, the graphs used make it difficult to observe the drawn conclusions (see technical corrections). If the aim is to validate the model results regarding the detection of MHWs, I suggest that what should be validated is the model's estimation of the various parameters characterizing MHWs (i.e., the number of events, maximum intensity, cumulative intensity, and MHW days). This comparison can be easily conducted from the data already calculated in Figure 4 and presented to the reader using scatter plots and linear regressions.

We will amend section 3.2.1 with a statistical evaluation of the agreement between the computed MHW metrics based on the observations and the model. The table below shows the Pearson correlation coefficients for the MHW metrics in Fig 4 between observational and model data for the two stations.

Table #: Pearson correlation coefficients from linear regression between the MHW metrics computed from the station data and the model data at the stations Lighthouse Kiel and Northern Baltic.

	common climatology period	MHW count	MH max intensity	MHW cumulative intensity	total MHW days
Lighthouse Kiel	1993-2021	0.82	0.88	0.66	0.93
Northern Baltic	1997-2021	0.74	0.89	0.82	0.94

5) The analysis of the vertical structure of MHWs is of great interest and scientific relevance. While I understand the space constraints inherent in contributions to OSR, I would encourage the authors to try to delve deeper into this analysis, as I believe it would enhance the scientific value of the contribution.
We will discuss Fig. 5 in more detail, especially regarding the subsurface MHWs and their potential drivers.

- **SPECIFIC COMMENTS**

L 10->The temperature anomalies are an intrinsic part of MHWs; I understand what the author means but “thermal anomalies”per se, are not a PREcondition for MHWs, actually they are a condition. Will be changed to “condition”

L 35->“In our BSH data”: Reword the sentence avoiding familiarity with the data used. Will be changed

L 41-49->As is the case with the other databases, this paragraph does not explain the purpose of reanalysis data in this work. Will be addressed

L 52->"The BACC Author Team (2008)": The word "The" is a definite article that, as far as I understand, does not take part in the name of the group. Thus, in the sentence, it must appear in lowercase, and the bibliographic reference must be “BACC Author Team” ordered under the letter 'B'. Will be changed

L 80-81->"rather continuously from 1989 until the present": Have MHWs been computed from incomplete datasets in this work? If so, discuss the implications.
Information on gaps in the observational data will be added to the text and uncertainties from this will be estimated based on Schlegel RW, Oliver ECJ, Hobday AJ and Smit AJ (2019) Detecting Marine Heatwaves With Sub-Optimal Data. *Front. Mar. Sci.* 6:737. doi: 10.3389/fmars.2019.00737

In the data from the two stations LT Kiel and Northern Baltic missing data has been quantified as follows:
LT Kiel 1993-2022: 8.6% of daily SST missing (9.1 % for 1989-2022)
Northern Baltic 1997-2022: 8.0 % of daily SST missing

According to Schlegel at al. (2019) for every percent point of missing data the degree of uncertainty introduced into the average marine heatwave (MHW) results is *on average* -0.44% in MHW count, -1.80% in duration and -0.31% in max. intensity.
Using a shorter time period than 30 years, e.g. here 1997-2022 for the station Northern Baltic, introduces an uncertainty of *on average* +0.32% per missing year on MHW count, -0.29% on duration and -0.23% on max. intensity. These average uncertainties can of course vary based on local conditions.

L 90->"1 nm": In scientific publications, units are expressed in accordance with the International System of Units (SI) or by using derived units as products of powers of these. Consequently, "1 nm"corresponds to one nanometre.I suggest expressing the spatial resolution in kilometres.
We will change the units to km

L 93->Section 2.4 employs the detection of MHWs, which is explained later in Section 2.5. Therefore, the description of methods for detecting MHWs must precede the model validation.
We can change the order

L 103-104->"In general though... station and model data": I think this is not well appreciated in the displayed figure. However, it would be interesting if, given that this Is a study of MHWs, a direct comparison of the parameters that typically characterize MHWs is made: Number of events, duration, intensity, etc. seeGeneral Comments). see point 4) above

L 105-114->Is it not possible to make a similar analysis for the deeper levels than the one performed in surface?
In terms of observational data we need stations with a long record of data. At Northern Baltic we only have SST and at Lighthouse Kiel the deepest datapoint is at 13 m, that is why we look at the model data in Fig 5 and section 3.2.2. We could only assess the deep 2022 MHW at Northern Baltic in a climatological context using model data. On the other hand, we do have limited space available for this report.
Please note in this context that there was a mistake in L. 82, where it was stated that the data at Northern Baltic was available down to depths of 103.8 m, which in fact is not the case. There is only surface data available. This will be corrected in the revised manuscript.

L 116->MHWs were previously defined. The definition will be removed

L 118-120->Despite of both packages produce identical results; it is worth mentioning why the effort of using different packages. Is it because of the computational efficiency of Matlab package? If so, why not use Matlab package also for observational data?
Out of the box, matlab can handle 2D data and we used it to produce 2D maps of 2022 MHW properties. The python code by default works with 1D data but also computes the block average for the climatological assessment of MHW.

L 126-129->In the work, the methodology for computing MHWs is applied not only to product ref.no. 3 but also to product ref. no. 2 (Figure 1 and 4). Therefore, both products deserve the same treatment here. If there were differences between the climatologies used for both products, Section 2.4 must include a discussion of how these differences can affect the validation.
We will harmonize which climatology period is used for all products and figures and will provide this in the revised manuscript. All MHW metrics will be computed based on the 1993-2021 climatology. Section 3.2 originally used 1993-2022.

L 144->SST anomaly rank is not a clear statistic. Please clarify.
The anomaly ranks shown in Fig. 2 are a very simple way to provide climatological context, providing information about how extreme an anomaly of a given magnitude is. For every grid point the anomalies of all years for a given month - e.g. August (calculated against the climatological mean of the respective month) - are sorted according to their magnitude. The anomaly rank depicts simply the rank in this sorting. For example, a plotted rank 1 of August 2022 in Fig. 2 therefore means that 2022 featured the hottest August in the entire dataset for the respective grid point. In Fig. 2 we show color shadings for the hottest eight and coldest eight ranks, respectively. We will provide a more detailed explanation in the revised manuscript.

L 154->See L 10. okay

L 165->“While the duration of …. Regions”: It is not clear what region you refer.
The whole paragraph will be rephrased. We will also add an additional table which will help improve the readability of this section.

L 175->“total five”: According to Fi. 4b there were four. A total of five in the observational data and four in the model data. This will be clarified in the revised manuscript.

L 178-179->“This trend of ….is taken into account”: This discussion is confusing. Why it is necessary to include extra data to detect trends? I would include the full observational record in left panels of Figure 4 avoiding this discussion.
The full observational record is added to Fig 4, left, Station LT Kiel.

L 195->“no temperature measurements exist in lower layers”: As far as I see, this is the first time the reader knows this lack of data in the text. This must be stated much earlier, in sections 2.2 and 2.4.
As mentioned, L. 82 falsely stated that there was measurement data available in lower levels at Northern Baltic. This will be corrected in the revised manuscript and we will clarify at this point that there are no long-term measurements in lower layers available that can be used to detect MHWs.

L 221-226->The information related to Holbrook et al. 2019 must be moved to Introduction. We will reorganize the discussion and move L. 217-226 to the introduction.

L 232->According to what I know, an exceedance of 9º C could be one of the highest intensities observed of a MHW in the world. It might be interesting to look for some bibliographic reference to provide some context of this huge magnitude.
As the MHW analysis will be redone using a climatological period from 1993 to 2021, this precise value may differ in the revised manuscript. Nevertheless, the maximum intensity here will remain rather high and we will discuss this in further detail, in particular with respect to local conditions (shallow water, sea ice, low seasonal temperature variations, etc.) that might lead to such enhanced local maximum intensities of MHWs.