

General comment

This manuscript describes an attempt to improve the methodology behind the Copernicus OMI indicator regarding SST extremes. The authors run an assessment of two different approaches regarding the use of different initialisation data (interim) or a multi-product ensemble approach that integrates regional and global model outputs with satellite-based observations. The main results show the benefits of these two approaches with respect to the actual procedure.

My only concern is that the analysis has only been run for 2024, which causes some uncertainty about the robustness of the new methodology. Indeed, the authors indicate in the conclusions section that a statistical analysis of the new methodology over decadal time scales would be necessary for a statistical characterisation of error distributions. This analysis would improve the consistency of the new approaches before operational applications.

As this seems to be the next step for the authors, my recommendation is to accept the manuscript for publication in its current form.

We thank the reviewer for their positive evaluation of the manuscript and for highlighting the limitation related to the temporal coverage of the analysis.

We fully agree that restricting the assessment to a single year (2024) limits the statistical robustness of the results. However, this constraint is largely inherent to the current operational configuration of the OMI SST extreme indicator, which relies on near-real-time (NRT) datasets that do not provide sufficiently long historical records to support medium- or long-term analyses.

As stated in the Conclusions, our intention is to address this limitation in future work by performing a decadal-scale assessment of the methodology. Although this may require adaptations to the original indicator configuration, relying exclusively on multi-year or reprocessed datasets, such an approach would enable a robust evaluation of long-term trends and interannual variability.

Minor comments

Which is the period referring to “the corresponding climatological 99th percentile”? Is it “(1993–2023)”? Why is this period used? Have different reference periods been tested?

The period should be indicated and justified in the data section, or different reference periods should be tested in the discussion sections.

We thank the reviewer for this comment. After a careful re-examination of the manuscript, we acknowledge that the definition of the reference period used to compute the climatological 99th percentile was insufficiently detailed in the methodology.

To address this, we have revised the methodological section and the relevant figure captions to explicitly include the following clarifications:

- 1. A justified explanation of the reference period used (1993–2023), which is primarily based on adopting a methodology as consistent as possible with the reference indicator under evaluation (OMI-NRT). Accordingly, the longest time period commonly available across all datasets considered is used as the reference climatology.*
- 2. A revision of all figure captions displaying anomalies to ensure that the reference period used for their computation is explicitly stated.*

Figure 3 caption does not describe panel (c)

We apologize for this oversight and commit to revising the caption of Figure 3 in the next version of the manuscript.

Manuscript revision

This study assesses two methodological improvements to the SST extremes Ocean Monitoring Indicator (OMI) in the IBI region: the replacement of the NRT dataset with the Interim product and the adoption of a multi-product ensemble approach. The Interim dataset improves internal consistency, whereas the ensemble method enhances robustness and allows the estimation of uncertainty. Overall, the work is relevant and well structured, but several major revisions are needed before publication.

1. Choice of the 99th-percentile anomaly as a metric

I have concerns regarding the use of the anomaly of the 99th percentile as the main metric for characterizing SST extremes, including the annual maps. The 99th percentile reflects only the very upper tail of the distribution and is therefore highly sensitive to short-lived spikes, noise, or inconsistencies across datasets—especially in coastal regions, where observational and modelling limitations are more pronounced. As a consequence, part of the signal may reflect artefacts rather than physically meaningful extreme events.

Although the 99th percentile captures the most intense warm anomalies, it may filter out persistent but moderate events that are ecologically relevant (e.g., typical marine heatwaves). Lower percentiles (e.g., 95th) or multi-metric approaches (e.g., combining intensity and duration) are commonly used in the literature and may offer more physically interpretable results.

Furthermore, because the current OMI is defined as the anomaly of the 99th percentile, the indicator becomes doubly sensitive—both the annual value and the baseline depend on extreme-tail behaviour. This amplifies uncertainty and complicates interpretation, especially when comparing the performance of different products or analysing trends.

We thank the reviewer for raising these important methodological considerations. We fully understand the concerns regarding the use of the 99th percentile anomaly and acknowledge that these points encourage future studies or modifications that could impact the definition of the indicator itself.

We note, however, that the indicator used in this study (OMI-NRT) is an established metric within the Copernicus Marine catalogue and has been referenced in previous Ocean State Reports published in this journal. A

detailed discussion on the choice of percentile is therefore more appropriately addressed within the context of these OSR contributions.

Including an in-depth evaluation of this aspect falls beyond the scope of the present work, which focuses on assessing the benefits of the multi-product methodology. For reasons of clarity and conciseness, such a discussion cannot be addressed in this study.

1. Methodological clarity on the annual anomaly maps

The manuscript does not clearly explain how the annual 99th-percentile anomaly maps are generated. While it is clear that the computation is performed on a per-pixel basis, the precise steps are not described. It remains unclear whether the annual 99th percentile is derived from daily SST values for the entire year, from a subset of data, or through some form of temporal aggregation or filtering. The construction of the climatological reference is also not explicitly documented—whether it is based on yearly percentiles averaged across years, or on other approaches.

Clarifying these aspects is crucial for transparency, reproducibility, and the physical interpretation of the resulting anomaly fields. As it stands, it is difficult to fully understand how Figures 1–2 were produced and how to be interpreted.

At this point, we identify two issues that we address separately: first, the lack of clarity regarding the methodology used to compute the annual 99th-percentile anomaly maps; and second, the need to explicitly describe the climatological reference period.

Regarding the clarity of the methodology used to compute the anomaly maps, we acknowledge that a more detailed explanation of the procedure for deriving the annual 99th-percentile anomaly maps is essential to facilitate understanding and ensure reproducibility. We therefore agree with the reviewer that this aspect should be clearly described. To address this issue, we propose to include in the Data and Methods section a step-by-step description of the methodology, including the mathematical formulation on which the OMI calculation is based, in order to provide readers with the necessary information to understand and reproduce the results.

Given that the indicator is computed using multi-year products (reanalysis datasets for model-based products and reprocessed datasets for satellite products), the first step of the methodology consists of computing the annual 99th percentile for each individual year of the multi-year product.

$$P_{99}^{MYP}(x, y) = \text{perc}_{99}(\{T^{MYP}(x, t) \mid t \in y\}) [1]$$

Where $T^{MYP}(x, t)$ is the SST field from the corresponding multi-year product at spatial location x and time t , and $perc_{99}$ denotes the 99th percentile computed over all temporal instants within year y .

These annual 99th percentiles are then used to estimate the mean annual 99th percentile, which serves as the reference climatology.

$$\bar{P}_{99}^{MYP}(x) = \frac{1}{N} \sum_{y=y_1}^{y_N} P_{99}^{MYP}(x, y) \quad [2]$$

In a second step, the 99th percentile for the year 2024 (y^{2024}) is computed using the corresponding NRT product.

$$P_{99}^{NRT}(x, y^{2024}) = perc_{99}(\{T^{NRT}(x, t) \mid t \in y^{2024}\}) \quad [3]$$

Where $T^{NRT}(x, t)$ is the field from the NRT product at spatial location x and time t including all available time steps within target year y^{2024} ,

Therefore, the annual anomaly map of the 99th percentile for the target year is finally obtained by subtracting the MYP mean annual 99th percentile from the NRT-derived 99th percentile.

$$A_{99}(x, y^{Yr}) = P_{99}^{NRT}(x, y^{Yr}) - \bar{P}_{99}^{MYP}(x) \quad [4]$$

Regarding the choice of the climatological reference period selected to compute $\bar{P}_{99}^{MYP}$, we acknowledge that this aspect was insufficiently detailed in the original manuscript. The reference period adopted in this study is 1993–2023, selected to ensure the highest possible consistency with the reference indicator under evaluation (OMI-NRT).

To address this, we have revised the methodological section and the relevant figure captions to explicitly include the following clarifications:

1. A justified explanation of the reference period used (1993–2023), which is primarily based on adopting a methodology as consistent as possible with the reference indicator under evaluation (OMI-NRT). Accordingly, the longest time period commonly available across all datasets considered is used as the reference climatology.
2. A revision of all figure captions displaying anomalies to ensure that the reference period used for their computation is explicitly stated.

1. Use of the terms “deterministic” and “probabilistic”

The manuscript frequently refers to “deterministic” and “probabilistic” methodologies. However, all three OMI formulations are based on statistical thresholds (percentiles), and the use of these adjectives may give the impression of fundamentally different methodological classes. I recommend removing these terms, unless a clearer conceptual distinction is provided and justified.

We acknowledge that the use of the terms “deterministic” and “probabilistic” may give the impression of fundamentally different methodological classes, which could be confusing without proper context.

In this work, the modification of the methodology through the application of multi-product strategies indeed represents a substantial evolution of the original indicator, affecting the type of results generated. While the original indicator provides a single, deterministic result (assumed to represent the “true” value), the use of ensemble-based methodologies leads to outputs that are distributed according to a probability function. In the context of ensemble systems, such outputs are commonly referred to as probabilistic results. Much of the interest of this study lies in assessing the reliability of this probability distribution.

Nonetheless, we fully understand the reviewer’s point that these terms should be carefully framed. We therefore consider it appropriate to revise the text to provide readers with the necessary context. In particular, Section 4 has been to explicitly introduce the conceptual distinction between the two frameworks at the point where the MULTI-OMI configuration is first discussed, clarifying that the shift from a single-product to a multi-product approach implies a transition from a single deterministic estimate to a range of plausible outcomes whose spread can be interpreted as a measure of uncertainty. In this revised version, the terms “deterministic” and “probabilistic” are retained but are now accompanied by the necessary contextual explanation to avoid misinterpretation, ensuring that the reader understands these adjectives refer to the nature of the output rather than to the underlying statistical methodology.

Section-specific comments

Abstract

- *Lines 17–19:* These lines repeat information already provided in lines 12–14; one of the two should be removed.

We agree with the reviewer. The paragraph will be revised, and the redundant information will be removed.

- *Line 14:* The expression “to support decision-making” is vague. Consider replacing it with “to characterize the reliability of this OMI” or something similar.

We will replace the expression with a more precise formulation, as suggested.

1. Introduction

- *Line 31:* The reference to “decision-makers and other stakeholders” appears unnecessary.

We will revise the sentence to remove the specific mention of “decision-makers and other stakeholders” as suggested. However, we consider that ocean indicators are a product that condenses information to facilitate the understanding of ocean state in a simplified way. Therefore, the sentence will be rewritten as follows:

“Ocean indicators can be defined as measures based on scientifically validated data and methodologies that enable the assessment of the state of ocean phenomena across various temporal and spatial scales, in a manner that is accessible and meaningful to end users.”

- *Line 33:* The citation “von Schuckman et al., 2026” appears incorrect; please verify (also at line 49).

We thank the reviewer for this comment. We will carefully check all occurrences of this reference in the manuscript to ensure they are correct. The correct citation is:

Von Schuckmann, K., Godoy-Faundez, A., Garçon, V., Muller-Karger, F. E., Evans, K., Appeltans, W., ... & Yu, W. (2026). Global ocean indicators: Marking pathways at the science-policy nexus. Marine Policy, 184, 106922, <https://doi.org/10.1016/j.marpol.2025.106922>.

- *Line 41:* Specify here the climatological period of the reanalysis product.

The climatological period of the reanalysis product will be included in the sentence.

- *Line 42:* The argument about “ensuring timely operational delivery” is unclear, since the OMI is updated annually (line 56).

We understand the reviewer’s confusion. By this statement, we refer to the inclusion of near-real-time (NRT) data, which allows the analysis of the most recent year to be produced rapidly, without waiting for the generation of multi-year products (MYP), which typically involve a delay of several months. In this sense, “timely operational delivery” refers to generating the analysis with the shortest possible latency. Nevertheless, we acknowledge that this expression may be confusing for the reader, and we will revise the text accordingly to improve clarity.

- *Lines 57–59:* This sentence is overly self-congratulatory; rephrase or support the claims with references.

The statement is derived from the analysis conducted in the paragraph in accordance with the framework proposed by von Schuckmann et al. (2026). Nevertheless, we acknowledge that the current wording may appear self-congratulatory. We will therefore revise the text to either provide a clearer and properly supported justification of this claim or remove the statement to avoid any unintended interpretation.

1. **Data and Methods**

- *Lines 91–92:* This repeats the idea introduced in line 46; consider removing.

This sentence will be removed.

- *Lines 122–127:* The construction of MULTI-OMI should be explained more thoroughly. Specify whether the ensemble anomaly is computed as a simple average, and confirm whether each product maintains its native spatial resolution (as suggested by lines 250–251). Also explain how non-significant pixels are defined.

The ensemble anomaly of MULTI-OMI is indeed the average of the anomalies obtained for its constituent members. These member anomalies are initially computed in the original resolution of their respective NRT products (every member is made up of a pair of one MY and one NRT product), but previously to the calculation of the ensemble average, all anomaly fields are interpolated using a conservative method to the IBI-NRT grid, which is the one in which OMI-NRT, OMI-INT and MULTI-OMI are intercompared. The ensemble anomaly is considered non-significant at those grid points where not every constituent member agrees on the sign of the anomaly. The text will be reworked to make these details clearer.

- *Line 129:* Consider replacing “common reference point” with “common reference dataset”.

We thank the suggestion and it will be included in the manuscript.

- *Lines 135–146:* This paragraph repeats earlier content and should be removed or rephrased.

This paragraph was intended as a concise summary to recapitulate, in a simplified manner, the detailed description provided in the section, thereby facilitating the overall flow and readability of the manuscript. However, in light of the reviewer’s suggestion, we will remove the paragraph to avoid redundancy.

1. Analysis

- *Line 194:* The expression “rather than a formal estimate of uncertainty” is confusing, since the ensemble is explicitly introduced to estimate uncertainty. Reword this sentence.

We agree that the expression may be confusing in this context, and we will remove “rather than as a formal estimate of uncertainty” to improve clarity.

- *Lines 195–196:* This sentence is too vague; rewrite or remove.

We agree that the sentence may be overly vague and will rewrite it to improve clarity and precision

- *Figure 2b:* The presence of STD values over non-significant areas (as defined in Fig. 2a) is confusing. Clarify how non-significant pixels are handled and be consistent with the definition (see also next comment).

We consider as non-significant areas all those grid points where not every member of the MULTI-OMI ensemble agrees on the sign of the anomaly. As we see it, having the same sign in all three members at a particular location supports the reliability of the ensemble anomaly’s positive or negative value, conferring robustness to the result in a qualitative sense. The STD of the ensemble, for its part, is considered in our contribution a measure of the quantitative agreement between the ensemble members, but we do not base the significance or non-significance of the ensemble mean anomaly on it (this possibility, however, was discussed and explored in the initial stages of our work). According to our chosen criterion, we consider significant, for example, that all three members of the ensemble have negative anomalies off the Portuguese coast (Fig. 2a), even though their particular intensities differ in some areas even in more than 1 °C (Fig. 2b). In this sense, the

ensemble STD can be seen as a complementary and specific indicator of the reliability of the MULTI-OMI's numerical value for the anomaly, and that is the reason why we display it separately from the non-significant areas. This explanation will be included in the manuscript to improve clarity for the reader.

1. Assessment

- *Line 238 / Fig. 3a:* Fig. 3a corresponds to Fig. 2a but without masking non-significant areas. It is not clear how the sign of anomalies is chosen in grey regions. In practice, it appears that in-situ observations are being overlaid onto areas classified as non-significant. Then, a consistent treatment of non-significant pixels is needed. Non-significant regions should be represented consistently (e.g., grey shading), otherwise interpretation becomes difficult. Considering the overall manuscript, I would suggest to adjust Fig. 3a to Fig. 2a accordingly.

We will revise Figure 3a to ensure a consistent treatment of non-significant regions, aligning it with the representation used in Figure 2a.

- *Lines 245–246:* The term “simplified metric” is unclear. If the metric cannot produce robust confidence intervals, its use should be justified.

We understand the concern and will revise this paragraph to better justify the use of a simplified metric and to clearly state the methodological limitations that prevent the calculation of statistically robust confidence intervals.

- *Figure 3b:* The meaning of the red and blue bars for OMI-OBS is unclear. Do they represent uncertainty? One would expect a single value, as for OMI-NRT and MULTI-OMI. Please revise. Moreover, OMI-INT is missing, and the caption lacks the description of Fig. 3c.

The red and blue bars corresponding to OMI-OBS in Figure 3b do not represent uncertainty. The OMI-OBS data are displayed in histogram format solely for visualization purposes, as this representation was considered clearer. The reference dataset (OMI-OBS) is shown using bars, while the compared results (OMI-NRT and MULTI-OMI) are represented using symbols. Nevertheless, we will revise the figure caption and the corresponding text to clearly explain what these bars represent.

Additionally, OMI-INT was not included in Figure 3b because the main objective of this work is to analyze the effect of the MULTI-OMI configuration relative to OMI-NRT. The use of OMI-INT is discussed in Section 5, where it is shown that its results are very similar to those obtained with OMI-NRT.

Therefore, the differences observed for MULTI-OMI cannot be solely explained by replacing the IBI-NRT dataset with IBI-INT. However, we understand the reviewer's confusion, as this point is not explicitly stated in the text before focusing exclusively on the MULTI-OMI results. We will therefore reorganize the structure of the text and Figure 3 so that the results of the Taylor diagram are first presented and discussed, leading to the conclusion that the subsequent analysis (Figure 3x and Table 2) will focus exclusively on the MULTI-OMI configuration.

- *Lines 254–255*: Provide a clearer explanation of why spatial variability is “more accurate”.

The statement that spatial variability is “more accurate” is a direct interpretation of the Taylor diagram, as it shows that the OMI-INT result lies closer to the line where:

$$\frac{\sigma_{mod}}{\sigma_{obs}} = 1$$

with σ_{mod} and σ_{obs} being the spatial standard deviations computed over all stations for model and observations respectively. Therefore, under equal correlation conditions, the model closest to this line provides a better representation of the spatial field. We will revise the text to improve clarity at this point, while avoiding a detailed explanation of Taylor diagram interpretation, which has been adequately described in previous studies (Taylor et al., 2001).

- *Line 263*: The sentence suggests that the decrease in STD for MULTI-OMI is problematic; clarify the intended meaning.

We understand that in this sentence the standard deviation of the sample data may be confused with the standard deviation of the ensemble members. Therefore, we will revise the sentence to provide a clearer explanation.

- *Line 265*: The RMSE how it is derived? From the Taylor diagram?

The RMSE referred to in the text is indeed derived from the Taylor diagram. Specifically, it is computed as the Euclidean distance between the point representing the model and the reference point representing the observations. We will revise the sentence to explicitly indicate this and improve clarity.

- *Table 2*: OMI-INT is missing and should be included.

OMI-INT was not included in Table 2 because the main objective of this work is to analyze the effect of the MULTI-OMI configuration relative to OMI-NRT. Moreover, as shown in Figure 3c, the results of OMI-NRT and OMI-INT are almost identical. We will review these data again, and if any relevant information is found, we commit to include it. However, if there are no significant differences compared to OMI-NRT, we do not see a clear motivation to include data that cannot be meaningfully discussed (see comment Figure 3b).

- *Lines 294–296:* MULTI-OMI does not reduce misclassification per se; rather, it provides the option to classify pixels as non-significant. Consider rephrasing accordingly.

We agree that the proposed wording reduces potential bias, and we will rephrase the sentence accordingly.

1. Conclusions

- *Lines 306–332:* This section repeats material already presented several times; a more concise synthesis is recommended.

While we appreciate starting the Conclusions with a brief summary of the work, we will follow the reviewer's suggestion and revise this section to provide a more concise synthesis.

- *Line 323:* “Deterministic error metrics” is not appropriate; mean bias and RMSE are standard statistical metrics.

We understand the point and will revise the sentence to refer to “error metrics of deterministic results” instead.

- *Line 329:* The claim that OMI-INT is more consistent because it derives from the same climate data record is not demonstrated. Provide justification or temper the statement.

The justification for this statement is described in the Methods section: “*This new dataset (referring to IBI-INT) is suitable for the calculation of the SST extremes indicator for two main reasons. First, it is generated using the same ocean model configuration as the IBI reanalysis (IBI-REA), which ensures greater homogeneity between the computational systems involved and reduces potential biases in the indicator’s estimation*” and “*...with the IBI-NRT dataset replaced by IBI-INT. This alternative formulation seeks to balance the trade-off between the update frequency of the indicator and the homogeneity of the upstream data used in its computation.*” Nevertheless, we will revise the sentence to make it clearer what is meant.

- *Line 380*: The term “substantial improvements” appears too strong given the results; consider softening the language. This applies elsewhere in the manuscript as well.

We thank the reviewer for this comment. We understand the suggestion and will revise the sentence to soften the language. Additionally, we will review the manuscript to avoid other statements that may appear too strong.