RC1

This study assessed the predictability of marine heatwaves surrounding New Caledonia, using dynamical coupled ocean-atmosphere model results. The assessment suggests that a probabilistic approach tends to have a higher skill than a deterministic approach, and stronger marine heatwaves are more predictable, as those that occurred during the La Niña events. The results suggest that the dynamical model predictions may provide useable forecast for regional stakeholders. The manuscript is generally well organised, and the following are some suggestions for the authors to improve their presentations.

We thank the reviewer for their positive feedback and suggestions, which we have included in the revised manuscript (see detailed responses below).

My main comment is about the model prediction skills. The authors admitted that "probabilistic and deterministic skill are impossible to compare directly since they are quantified with different scores". So the conclusion that "the probabilistic approach improves the quality of the forecast" is not very intuitive. Maybe 1-2 case studies on some of the recent marine heatwave event predictions can better demonstrate this point. Such as, the statistical prediction provides a more advanced warning for stakeholders.

We agree that the arguments supporting this key result weren't presented clearly enough. We based this assessment on two criteria : the comparison with the persistence forecast and the comparison with estimates of "no skill" and "good" scores from the literature. Horizontal lines drawn in the figures, showing the "good" threshold were already included and we have added lines for the "no skill" threshold in the revised manuscript. We have rewritten the relevant paragraph in the Results section to emphasize these points and illustrated it with the example shown in Fig 1 (modified sentences are in bold):

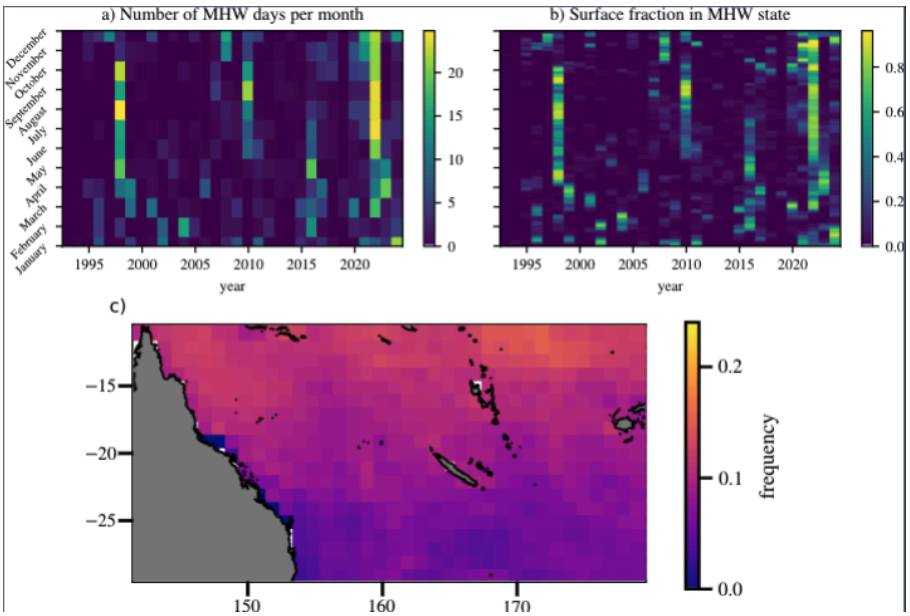
Both forecast systems perform well with the probabilistic approach, although the skill of ECMWF (Fig. 7) is again higher than that of Météo-France (Fig. 6). ***For the most intense events in particular (dark purple lines in Figures 7-8), the*** *Area Under the Curve (****AUC) remains above 0.8 for at least 4 months in the Météo-France forecasts and 6 months in the ECMWF forecasts. In addition, probabilistic forecasts tend to be better than persistence, even at very short leadtimes, while the deterministic forecasts only start to outperform the persistence after the first few weeks or even months (bold lines in Figures 4-5-6-7).*** *While probabilistic and deterministic skills are impossible to compare directly since they are quantified with different scores (the AUC and the Pearson correlation coefficient, respectively), these results suggest that the probabilistic approach improves the quality of the forecast.* ***Indeed, forecasts are generally considered to have no skill when the Pearson coefficient and the AUC are below 0.5, and to have "good" skill when the Pearson coefficient is above 0.7 or the AUC is above 0.8 (Horizontal dashed lines in Figs 4-5-6-7). Even if these reference values for what constitutes "good skill" are subjective and could be debated , there is a clear contrast between the deterministic and probabilistic scores: deterministic scores fall under these thresholds much quicker than probabilistic scores. For instance, all AUC values are above 0.5 for the entire forecast (7 months), while only some metrics have a Pearson coefficient above 0.5. Pearson coefficients only exceed 0.7 at very short leadtimes, while AUC commonly exceeds 0.8 for many months. As an illustration, Fig. 1 shows how a MHW in february 2016 can be detected 1 month in advance by a probabilistic forecast (red line) but completely missed by a deterministic forecast (black line).***

Is this conclusion sensitive to the selection of the 20% criteria? The authors may want to provide more details about the AUC calculation in the supporting information.

To clarify the relationship between the AUC (and the conclusions based on it) and the choice of decision threshold, we have rephrased the explanation of the computation of the AUC in the main text as follows : "***The Area Under the Curve (AUC) is computed as the integral of the ROC curve : it varies between 0.5 (when the ROC curve is the diagonal) and 1 (when the ROC curve reaches the top left corner). Since the AUC is constructed using a range of decision thresholds, it is not sensitive to the choice of a particular threshold; rather it allows an operator to choose a threshold, which compromises between the number of true positive and false alarms depending on what the operator decides to be most important for his decision making (see the Discussion section). The AUC is a commonly used measure of skill, with a score of 0.8 generally considered to represent a useful forecast \citep{Corbacoglu2023}. The AUC measures the discrimination, i.e. the ability of the forecast to distinguish between events and nonevents \citep{BenBouallegue2022}."***

Figure 3 a and b: it is better to use a Hofmuller diagram to show the year to year variations.

Thank you for the suggestion. We have modified the figure and believe it is now much easier to read.



RC2

This manuscript makes a meaningful contribution to the growing field of Marine Heatwave (MHW) forecasting, specifically addressing a gap in regional scale predictability assessments for South Pacific island nations. The focus on New Caledonia as a case study, along with explicit attention to forecast usefulness for local stakeholders, adds practical value beyond typical global skill assessments. The finding that cold season MHWs during La Niña events are predictable up to 7 months in advance, while warm season MHWs have little skill beyond a few weeks, is a substantive and actionable result. The systematic comparison of deterministic versus probabilistic approaches across multiple MHW metrics is also a useful contribution. It is also worth acknowledging that forecasting MHWs at seasonal timescales in a dynamically complex region like the Southwest Pacific is genuinely difficult, and the paper's honest treatment of skill limitations, particularly the consistent reporting of results with and without the dominant La Niña years, is commendable.

However, I have a few concerns.

We thank the reviewer for their positive feedback and we hope to have addressed their concerns in the revised manuscript (see detailed responses below).

1. The verification framework is somewhat incomplete. The study evaluates discrimination using AUC and correlation, but does not assess calibration. From a user perspective, this is important. It is one thing for a forecast to correctly rank higher risk periods, but another for the probabilities to be reliable. For example, if a forecast gives a 20 percent probability of a MHW, it would be useful to know whether such events actually occur around 20 percent of the time. Reliability diagrams or similar diagnostics are standard in ensemble forecast evaluation and would make the probabilistic results more interpretable.

We agree with the reviewer. We have added reliability diagrams as supplementary material and they are referenced in the text as follows :

- in the methods/verification section : ***In addition, the calibration of the forecast was evaluated using reliability diagrams (Supp. Fig. A10-13)***
- in the results section : ***Reliability diagrams show that the forecasts are well-calibrated, although both systems tend to overpredict MHWs (Supp Figs A10-13)***
- in the figure legends : ***Fig A10-11 : Reliability diagram (blue line) for the two severity levels of exposure : low (top row, 1 ≤ severity ≤ 1.5) and high (bottom row, (severity ≥ 1.5). In each probability bin (x-axis), the blue line shows the observed frequency among all days where the forecast predicted this probability. The diagonal represents a perfectly calibrated forecast. Since the blue line is below the diagonal, the diagram shows that the forecast (here Météo-France System 8) usually overpredicts MHW events. The green line shows the number of samples in each probability bin.***
  Fig A12-A13 : Same as Fig A10-11, but for the ECMWF SEAS5 system. The forecast is better calibrated than Météo-France System 8, but also overpredicts MHW events.

We also provide the climatological frequencies of all the MHW events considered as context for the interpretation of the probabilistic forecasts, as we believe they may be more intuitive than reliability diagrams to some users (Tables A1-4).

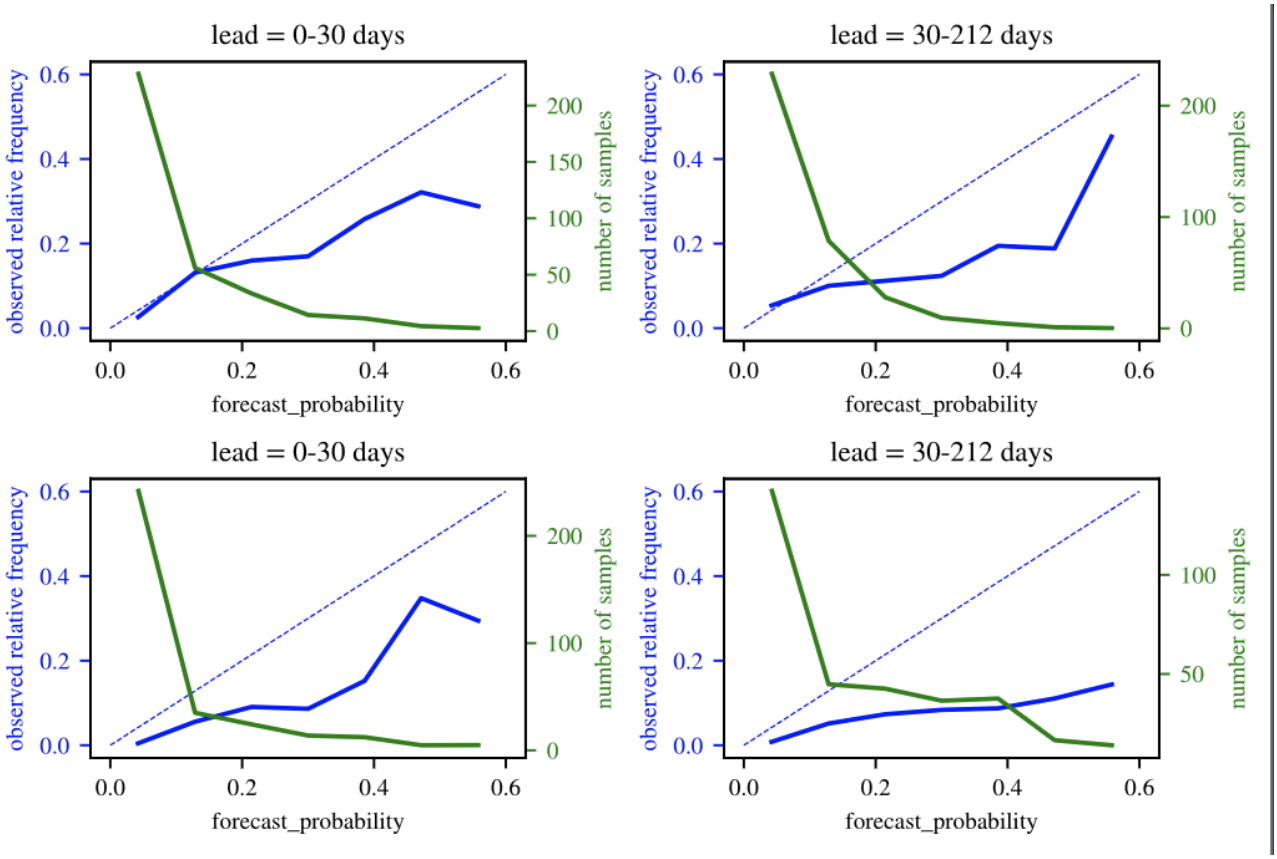


Fig A10

2. The 20 percent decision threshold is also selected using the same hindcast dataset. As a practical recommendation, a simple cross validation approach such as leave-one-year-out would help confirm that this threshold holds up out of sample, especially given the limited sample size when stratifying by season or ENSO phase.

We thank the reviewer for this remark, however we have decided not to perform a rigorous cross-validation because none of our conclusions are sensitive to the choice of threshold : it is included only in the Discussion section, as a starting point for co-constructing bulletins with local stakeholders (see our answer to the 3rd comment below). In fact, the comparison of the full hindcast and the hindcast with the three years removed constitutes a cursory cross-validation. The robustness of the threshold is also supported by the comparison between the Meteo-France and ECMWF hindcasts, and the expansion to a wider southwest Pacific region. In all cases the 20% appears as a reasonable choice of decision threshold. In addition, this choice is in part motivated by its intuitiveness to the stakeholders : 20% is a round number and corresponds to a doubling of the climatological probability of a MHW (10%, as defined by the 90th percentile). This is also why we have decided to recommend a single value for all metrics and leadtimes, even though it will perform better for some than others.

We have expanded on this last point in the Discussion section : *Second,* ***we recommend raising a "MHW alert" when an optimal decision threshold of 20% is reached (i.e., when more than 20% of the members predict a MHW) in order to maximize hits and minimize false alarms. This choice is supported by the shape of the ROC curves (Fig. \ref{probabilistic_skill}), and is also intuitive : 20% is a round number and corresponds to a doubling of the climatological probability of a MHW (which the 90th percentile constrains at 10%). The climatological probabilities of all the MHW events considered (for instance a MHW of***

category 2, or covering more than 50% of the EEZ) needs to be made available along with the forecasts to provide context to guide their interpretation (Tables A1-4). Also, for simplicity, we recommend a single value for all metrics and leadtimes in both forecast systems, even though the 20% threshold will perform better in some cases than others. *In practice, this means…*

3. Also, while the paper makes a good case that the forecasts are useful, it only partially addresses usability and does not fully engage with the third step of the Spillman et al. (2025) framework, which is ensuring that forecasts are actually used. As a user, the paper provides a clear sense of when the forecasts are likely to have skill, but less clarity on how they would be applied in practice. For example, there is limited guidance on how to access or process the forecast data, compute the metrics operationally, interpret ENSO phase in real time, or translate a 20 percent ensemble threshold into a concrete management decision.

As the reviewer noted, this paper is only the first phase in the process of producing an operational forecast. The results presented here are not standalone, but are part of a larger interdisciplinary project (www.MaHeWa.fr), which pursues many objectives that are beyond the scope of the present study, but will eventually fully address the concerns of the reviewer.
Future perspectives and on-going work include:

- investigating other forecasting systems (shorter leadtimes at 10, 28 days, but also 7 months using a higher resolution model, and also a machine learning approach)
- working with the operational center in oceanography Mercator-Ocean, to produce forecasts operationally and compute the metrics for online availability.
- we organized two transdisciplinary workshops with scientists and stakeholders to co-design the forecast bulletins in order to maximize stakeholder engagement.
- For an efficient communication, the forecasts bulletins will be distributed through the French National Met. Office: “Meteo France”. A communication program is also put in place, to ensure that the information will be transmitted first to the most relevant decision-makers.

To clarify this important point, we have added the following paragraph to the Conclusion :
Future work will increase the usefulness, useability and use of the forecasts. Social science studies can contribute to the usefulness of the forecasts by identifying the vulnerabilities of local communities to MHWs, and the spatio-temporal scales where forecast information may be actionable. Different forecasting systems, such as a higher resolution model or machine learning approaches, may provide higher skill and thus improve useability. In order to maximize stakeholder engagement, the bulletins are co-designed using trans-disciplinary approaches. Close collaborations between academics and operational centers (Mercator Ocean International and Météo-France) prove to be essential in order to efficiently produce and distribute appropriate forecasts.

Overall, this is a solid and useful study. Addressing these points would strengthen both the scientific interpretation and the practical relevance of the work.