Reviewer #1's comment:

In this manuscript, the authors assess the potential improvements or added value of MED-CORDEX RCSM relative to the GCM driving the RCSM downscaled results. The authors use a set of models coming from CMIP5 and CMIP6.

In the different sections of the manuscript, the authors run a complete set of statistical and spatial analysis of different SST and MHW metrics. This is a very comprehensive study, as it carries out an exhaustive analysis of all the models and the multi-model approach for various relevant metrics. The analyses are well-founded and describe the changes observed in each of the models, explaining the improvement or deterioration in the results in each case. It is interesting that the authors discuss both cases where the impact of using an RCSM is positive and those where it is negative. Furthermore, in some cases, possible causes of the change—whichever direction it takes—are indicated, depending on the models used.

The main conclusion from this manuscript is that the use of MED-CORDEX models, with higher spatial resolution, do not necessarily yield an improvement of SST and MHW analysis and metrics characterization. Whether the results show an improvement or a deterioration depends both on the metric analysed and the model used, as well as its specific configuration. Consequently, it cannot be stated that the higher spatial resolution of RCSM models leads to a general improvement in results. Although this may, on the face of it, seem a discouraging finding, it enables researchers in the field of modelling to identify areas for improvement in future generations of models. In this way, improvements can be incorporated into those model configurations that are yielding poorer results, whilst reinforcing those aspects or processes linked to their improvement.

Hence, my recommendation is to publish the article in its present form by clearly stating that it is an assessment that could improve the future generation of models. The authors should also address the questions below.

QUESTIONS:

CMIP5 and CMIP6 models are analysed together, is there an impact on the results because of comparing models coming from different CMIPs?

For the comparison of the GCM vs RCSM, the authors calculates metrics and statistics in coincident grid points and exclude "empty" points when only one of the GCM or RCSM has data. Have the authors analysed if there is a bias in the results because of excluding points with valid data? How many points are excluded? Is it only a small subset of data points?

Figure 1 (and others) shaded colours do not look like green, red, orange in my pdf.

The authors describe improvement in coastal areas for a good set of metrics and models. Usually coastal areas are "problematic" from the observational point of view at least from satellites. Is there the same problem in the oceanic model as the

coastal areas are the model boundaries? Do the authors have a hypothesis why is there more improvement in the coastal areas and semi-enclosed seas?
Answer:

We sincerely thank the reviewer for reading our manuscript and for the constructive and positive assessment. We are pleased that the overall quality and scope of the analysis were found to be solid. Below we address each point raised by the reviewer. In the following, the points and questions raised by the reviewer are reproduced in bold, text from the first version of the manuscript is shown in red, and planned changes and additions are shown in blue. Planned changes and additions will be implemented in the revised version of the manuscript, which will be submitted after the closing of the public discussion. All lines refer to the first version of the manuscript.

Main point: "Hence, my recommendation is to publish the article in its present form by clearly stating that it is an assessment that could improve the future generation of models."

We note that this framing is already present in the manuscript, in particular in the Introduction (Section 1, lines 60-61):

"This assessment evaluates the effectiveness of downscaling for marine extremes and informs future modeling priorities."

However, we agree it is not sufficiently clear in the Discussion. We thus plan to add the following sentence at the end of the first paragraph of Section 4:

"Identifying where and why regional downscaling succeeds or fails represents a step towards improving future generations of regional climate models."

Question 1 — "CMIP5 and CMIP6 models are analysed together, is there an impact on the results because of comparing models coming from different CMIPs?"

As already discussed in the manuscript (Section 3.2.1, lines 252-253), the two CMIP6-driven configurations (GUF and CNRM-RCSM6) are forced by GCMs that more accurately capture mean MHW intensity compared to the CMIP5-driven models. We further detail this statement by recognizing that it's a result already present in the literature:

"Interestingly, the two CMIP6 GCMs accurately capture the mean MHW intensity, deviating from the general CMIP5 behaviour, consistent with the improved MHW skill of CMIP6 models reported in the literature (Qiu et al., 2021)."

Lower mean biases in CMIP6 GCMs may limit the room for improvement by the corresponding RCSMs. However, we find that Added Value in MHW intensity is related more to model-specific factors. A similar argument, already present in the manuscript, applies to MHW duration: models driven by GCMs with a larger bias (predominantly CMIP5) tend to exhibit larger positive AVi, while those driven by more accurate GCMs (CMIP6) show more limited improvement (Section 3.2.2, lines 272-275):

"Notably, the ability of RCSMs to improve MHW duration appears partly conditioned by the bias of their driving GCM: models driven by GCMs with a large bias (CMCC-CM, MPI-ESM-LR and IPSL-CM5A-MR, Fig. S13d,h,l) tend to exhibit larger positive AVi, while those driven by GCMs with a small bias (CNRM-CM5 and CNRM-ESM2-1, Fig. S13f,j,r) show more limited improvement."

We plan to make these considerations more explicit in Section 4 of the revised manuscript, where we want to acknowledge that CMIP6 GCMs tend to be more accurate for MHW metrics and this has implications for the quantification of the added value of the downscaling models:
"It is worth noting that CMIP6 GCMs tend to exhibit lower mean biases for both MHW intensity and duration relative to CMIP5 GCMs, which naturally reduces the room for improvement by the corresponding RCSMs in terms of mean bias correction. The reduced magnitude of the driving GCM bias partly conditions the RCSM ability to show added value in MHW duration, as discussed in Section 3.2.2, while no such systematic relation is found for MHW intensity, where added value is governed by model-specific configuration factors."

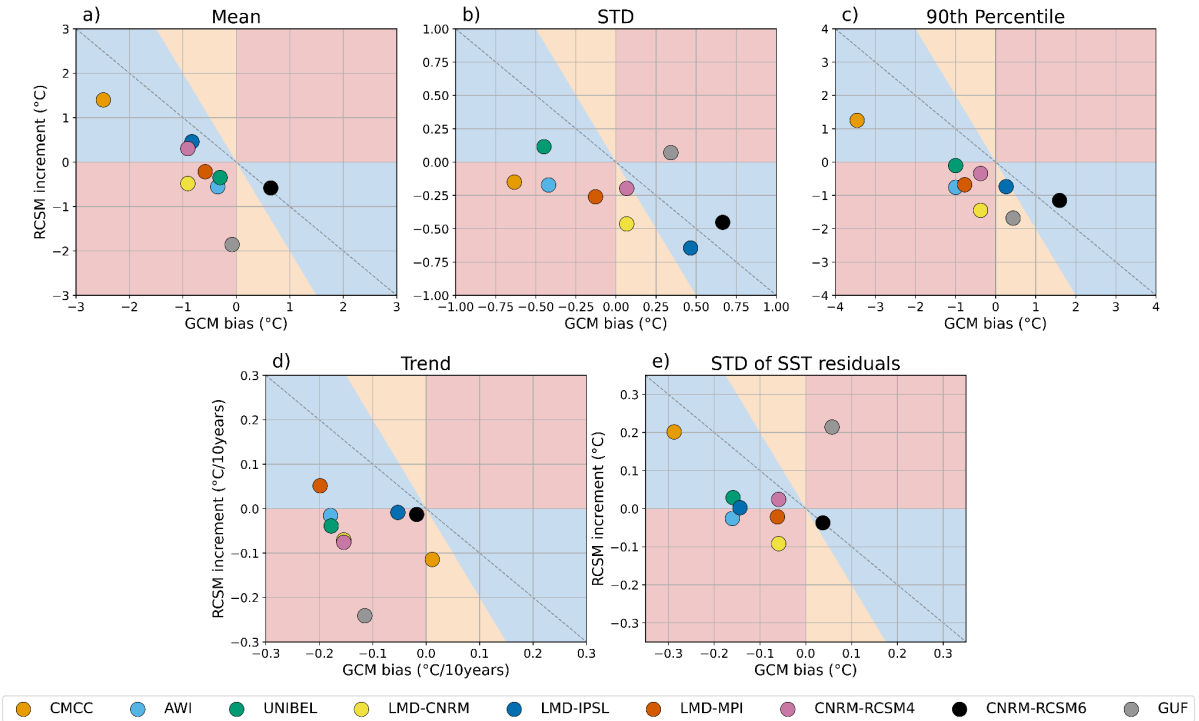
Question 2 — For the comparison of the GCM vs RCSM, the authors calculate metrics and statistics in coincident grid points and exclude 'empty' points when only one of the GCM or RCSM has data. Have the authors analysed if there is a bias in the results because of excluding points with valid data?

We thank the reviewer for raising this point, though we would like to clarify how the spatial comparison is conducted in our study. We do not exclude any grid points from the analysis. Rather, for each model family, observations and GCM outputs are bilinearly regridded onto the native RCSM grid, so that all three datasets (OBS, GCM, RCSM) are compared at exactly the same grid points, i.e., those of the RCSM. This is standard practice in added value studies and ensures that the comparison is spatially consistent without any data exclusion. Lines 121-126 describe the adopted methodology:
"For each model family, OBS and GCM were bilinearly regridded on the correspondent RCSM grid to facilitate a fair comparison between regional and global models. Interpolating GCM values to a finer resolution may produce unrealistic outputs, while interpolating OBS and RCSM values to a coarser resolution can compromise the representation of the PDF (Prein et al., 2016). However, to effectively illustrate the added value of downscaling, it is crucial to demonstrate that

resolving local-scale processes provides more information than simply interpolating a coarse, non-resolving grid onto a finer grid."

If the reviewer is referring to the "empty" grid points shown in Figure 2 and similar figures, these are locations where AVi does not significantly differ from the AVi calculated from a random distribution, i.e., the AVi is neither positive nor negative. For further details, refer to Section 2.2.2, lines 115-119:
"At each grid point, the statistical significance of AVi was assessed via bootstrap randomization. A reference distribution of 1000 synthetic AVi values was generated by randomly drawing two new GCM and RCSM samples from their union, simulating the null hypothesis of no systematic difference between global and regional models. The original AVi was defined significant when it fell outside the 5th–95th percentile range of this distribution. The area-averaged AVi was then computed over the Mediterranean domain, with non-significant grid points set to zero."

Question 3 — Figure 1 (and others) shaded colours do not look like green, red, orange in my pdf.
We thank the reviewer for raising this point. Following a similar comment from Reviewer 2, we took the opportunity to revise the color scheme of Figures 1, 4, S4, and S12 to a colorblind-friendly palette, which also improves overall clarity and contrast.
The following is the updated version of Figure 1 for reference.



We will update the caption as follows:

“Figure 1. RCSM increment versus GCM bias for area-averaged SST (a) mean, (b) STD, (c) 90th percentile, (d) trend, (e) STD of SST residuals over 1982-2020. The grey line indicates perfect error correction. Blue, red, and orange shading denote skillful error reduction, deterioration, and overcompensation (RCSM error opposite in sign but larger in magnitude than the GCM error), respectively. For the definition of the Mediterranean domain see Fig. 2.”

Question 4 — "The authors describe improvement in coastal areas for a good set of metrics and models. Usually coastal areas are 'problematic' from the observational point of view at least from satellites. Is there the same problem in the oceanic model as the coastal areas are the model boundaries? Do the authors have a hypothesis why is there more improvement in the coastal areas and semi-enclosed seas?"

The consistent improvement found in coastal areas and semi-enclosed seas is indeed one finding of our study, and we believe it reflects an added value of regional downscaling rather than an observational or model boundary artifact. The Med-CORDEX RCSMs considered here are specifically designed with enhanced horizontal resolution in the ocean (6-30 km vs. ~100 km in the driving GCMs), which allows them to better resolve coastline geometry, complex topography, strait dynamics, and mesoscale circulation features. Additionally, the higher atmospheric resolution and increased coupling frequency of the RCSMs improve the representation of regional wind systems and heat fluxes that strongly modulate SST in coastal regions. Together, these structural improvements provide a physically consistent explanation for the systematic added value found along the coastlines and in semi-enclosed areas, consistent with what is already discussed in the manuscript. The skillful bias reduction in coastal areas and semi-enclosed seas is highlighted in the Results (Section 3.1.1):

(lines 167–169)
“Yet, all RCSMs show skillful bias reduction in key regions, particularly along the coastlines and in semi-enclosed areas such as the Adriatic Sea, the Alboran Sea, and the Aegean Sea, where the added value of downscaling is most evident (Fig. S3).”

(lines 174-176)
“As with the mean bias, RCSMs tend to degrade the STD over large portions of the basin, while showing localized improvement along the coastlines and in semi-enclosed areas (Fig. S5).”

(lines 192-193)
“As noted for the mean bias and STD, spatial improvement is again concentrated along the coastlines and in semi-enclosed areas (Fig. S8).”

The physical mechanisms linking higher resolution to improved mesoscale dynamics and wind-driven forcing are discussed in the Results (Section 3.2.2), specifically in relation to MHW duration:

(lines 276–286)
“This result reflects the broad improvement in SST variability discussed above and agrees with prior findings (Pilo et al., 2019; Guo et al., 2019; Capotondi et al., 2024), showing that improved performance in MHW duration can be achieved by enhancing horizontal resolution, which plays a key role in better resolving the temporal build-up and dissipation of MHW events through enhanced mesoscale dynamics. RCSMs increase the realism of local and regional dynamics, producing SST fields with sharper gradients and more accurate spatial variability that better capture the conditions leading to the onset and decay of short-lived events. In the Mediterranean, MHW termination is frequently driven by regional wind systems which trigger rapid surface cooling through enhanced heat loss (Bonino et al., 2025). These mesoscale wind patterns are inherently better captured at higher resolution (Herrmann et al., 2011; Obermann et al., 2018), enabling RCSMs to more realistically simulate the atmospheric forcing episodes responsible for MHW decay. Beyond horizontal resolution, air–sea flux parameterizations and the thickness of the uppermost ocean layer further modulate the ocean’s response time to such forcing, and are therefore also expected to influence MHW duration.”

The Discussion (Section 4) already summarizes that the added value of dynamical downscaling is concentrated where fine-scale ocean-atmosphere interactions and topographic constraints are most influential:

(lines 358–362)
“Nevertheless, 7 out of 9 RCSMs improve the spatial structure of the mean SST field relative to their driving GCMs and all RCSMs show error improvement localized along the coastlines and in semi-enclosed areas such as the Adriatic, Alboran, and Aegean Seas, regardless of their basin-averaged performance. Together, these results confirm that the added value of dynamical downscaling for SST is primarily a regional-scale phenomenon, concentrated where fine-scale ocean-atmosphere interactions and topographic constraints are most influential.”

We plan to further clarify this aspect by adding a sentence in the Abstract:

“SST improvements are most consistent along coastlines and in semi-enclosed areas, where fine-scale ocean-atmosphere interactions and topographic constraints are best resolved by RCSMs.”

Reviewer #2's comment:

review of "The added value of Med-CORDEX Coupled High-Resolution Regional Climate Models in representing Sea Surface Temperature and Marine Heatwaves in the Mediterranean Sea" by De Rovere et al

This study applies a series of diagnostics to the daily SST field from ~10 MedCordex simulations and their parent GCM, so as to diagnose when and where RCSMs adds value. The key result is that except for MHW duration and certain aspects of the PDF of SST variability where RCSMs clearly adds value, there is little evidence for systematic improvements across models away from coastlines or regions with topography. Further, rather than horizontal resolution adding substantial value, the authors seem to highlight vertical resolution in the mixed layer as a potentially more important model setting, though in an ensemble of opportunity as used here it is difficult to isolate processes.

This is a solid study that is nearly ready to be accepted. It also highlights the limitations of dynamical downscaling- this reviewer would have hoped that all of this CPU time and effort would have helped matters more. But there is one major issue: I expect a more careful assessment of the pros and cons of the diagnostics used. Specifically, they aren't using process-based diagnostics that attempt to follow the processes that drive SST variability, rather just examine SST itself. For example, we know that MHWs (and SST variability more generally) are driven by anomalies in evaporation, which in turn is driven by weak wind speed and enhanced atmospheric near-surface humidity, and also by enhanced surface downwelling radiation associated with fewer clouds (Bonino et al, already cited; Garfinkel et al 2026; GRL). I think the assessment of SST variability and MHWs might be more informative if the authors analyzed how these processes are represented in GCM vs. RCSMs. I would expect that evaporation would be better handled with enhanced horizontal resolution as fronts are now resolved, and the representation of clouds will also be very different in a GCM vs. RCSMs especially if the RCSM turns off convection. But this needs to examined closely. The paper currently discussed how oceanic processes might differ in a GCM vs. a RCSMs, but atmospheric processes matter too (Bonino et al and Garfinkel et al 2026) and they are completely ignored.

To be specific, I don't think the current study needs to include the analysis I'm suggesting above. Rather, a new paragraph in the Discussion would suffice. This is mainly because at least four additional fields would be needed - near-surface wind speed, surface downwelling shortwave, cloud fraction, specific humidity in the atmospheric surface layer - and I don't know if this data is even available. But I certainly think that more could be done with the GCM and RCSMs used here to better understand why RCSMs don't help as much as this reviewer would have hoped. If this data is not available, then this could be a recommendation for the data request in future phases of the Cordex project. More generally, I suggest adding a sentence to the Discussion that notes the limitations of dynamical downscaling; this reviewer would have hoped that all of this CPU time and effort would help matters more.

specific comments:

1. Abstract: The role of vertical resolution in the upper ocean seems to be more important than that of horizontal resolution (based on the Discussion). This would seem to me to be an important point worth highlighting in the abstract.

2 line 82: "We focus on such period to maximizes overlap with" needs to be rewritten

3. Figure 1 and similar: the red and green shading are not colorblind friendly. Please consider changing these colors to ones with more contrast. Also, why doesn't the orange overcompensation shading extend all the way to the green line for perfect error correction?

line 179, 202, 204, 229,230,272: My first language isn't English either, but I think "impoverished" is used incorrectly. Do you mean degraded?

Answer:

We sincerely thank the reviewer for the careful reading of our manuscript and for the constructive comments, which we believe have improved the quality of the paper. We are pleased that the overall quality of the study was found to be solid. Below we address each point raised by the reviewer.
In the following, the points and questions raised by the reviewer are reproduced in bold, text from the first version of the manuscript is shown in red, and planned changes and additions are shown in blue. Planned changes and additions will be implemented in the revised version of the manuscript, which will be submitted after the closing of the public discussion. All lines refer to the first version of the manuscript.

Main point: "The authors currently discussed how oceanic processes might differ in a GCM vs. a RCSMs, but atmospheric processes matter too *(Bonino et al., 2025; and Garfinkel et al. 2026)* and they are completely ignored."

We thank the reviewer for this insightful comment and for the suggested reference. We agree that a process-based assessment of how atmospheric drivers of MHWs, in particular wind speed, evaporation, cloud fraction, and near-surface humidity, are represented across the Med-CORDEX ensemble would provide a more complete picture of where and why RCSMs add value. We acknowledge this as a limitation of the current study. We plan to add a new paragraph in the Discussion (in between lines 399 and 400) explicitly acknowledging this limitation and recommending it as a priority for future work:

"The present study focuses exclusively on SST and derived MHW properties, without examining the atmospheric drivers that govern SST variability and MHW onset. Recent work has shown that Mediterranean MHWs are primarily driven by anomalies in surface heat fluxes, in particular reduced latent heat loss associated with weak winds, enhanced near-surface humidity, and reduced cloud cover associated with enhanced surface downwelling radiation (Bonino et al., 2025; Garfinkel et al., 2026). RCSMs improve the representation of atmospheric fronts and mesoscale wind systems (Herrmann et al., 2011; Obermann et al., 2018), which is expected to translate into better representation of surface latent heat fluxes and evaporation (Hamon et al., 2016; Herrman et al., 2016), key drivers of MHW onset and decay. However, a process-based assessment of how such atmospheric processes influence MHWs across the Med-CORDEX ensemble was beyond the scope of this study but represents an important direction for future work."

Specific comments:

1. Abstract: The role of vertical resolution in the upper ocean seems to be more important than that of horizontal resolution (based on the Discussion). This would seem to me to be an important point worth highlighting in the abstract.

Thank you for pointing this out. We plan to add the following sentence at the end of the abstract:
“In particular, within the Med-CORDEX ensemble, inter-model differences in MHW added value appear to be correlated with improvements in vertical resolution.”

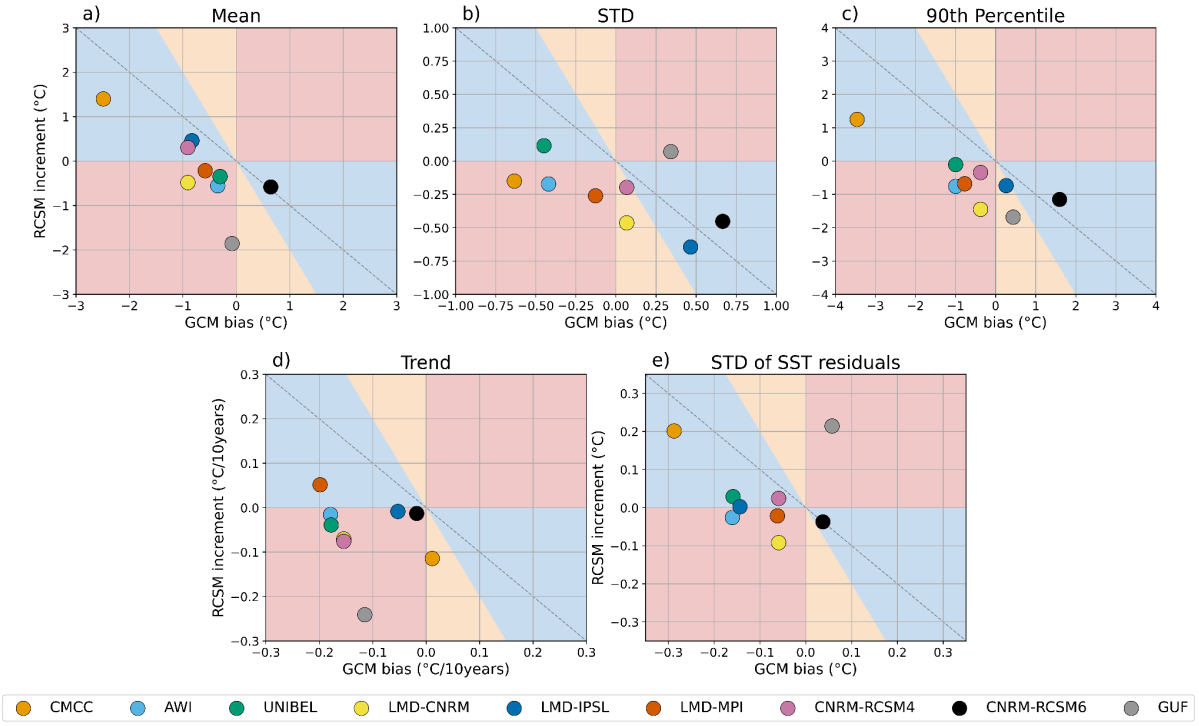
2. line 82: "We focus on such period to maximizes overlap with" needs to be rewritten

We will rewrite the sentence to:
“We select this period to maximize overlap with high-resolution OBS, improving the statistical robustness of the results while minimizing the use of data from future scenario simulations.”

3. Figure 1 and similar: the red and green shading are not colorblind friendly. Please consider changing these colors to ones with more contrast. Also, why doesn't the orange overcompensation shading extend all the way to the green line for perfect error correction?

Thank you for raising this point. We have modified all scatterplots to use colorblind-friendly colors for both the background shading and the dots in Figures 1, 4, S4, and S12.

The following is the updated version of Figure 1 for reference.



We will update the caption as follows:
"Figure 1. RCSM increment versus GCM bias for area-averaged SST (a) mean, (b) STD, (c) 90th percentile, (d) trend, (e) STD of SST residuals over 1982-2020. The grey line indicates perfect error correction. Blue, red, and orange shading denote skillful error reduction, deterioration, and overcompensation (RCSM error opposite in sign but larger in magnitude than the GCM error), respectively. For the definition of the Mediterranean domain see Fig. 2."

Regarding the orange overcompensation shading, this represents the area where the RCSM bias is larger in magnitude than the GCM bias but opposite in sign. This region is bounded by the lines RCSMincrement = −0.5 × GCMbias and GCMbias = 0. In the new figures, the blue region extending from the orange shading to the line of perfect error correction identifies the area where the RCSM bias has opposite sign compared to the GCM bias but is smaller in magnitude.

line 179, 202, 204, 229,230,272: My first language isn't English either, but I think "impoverished" is used incorrectly. Do you mean degraded?

Yes, we will replace all "impoverished/impoverishment" entries with "degraded/degradation". Thank you for the suggestion.