Review for Bossert et al. “Ocean and marine heatwaves responses to multiple net-zero worlds”

Summary of the manuscript:

In this study, the authors investigate the zero-emission commitment (ZEC) of ocean temperature changes across a range of different global warming levels (GWLs, from 1.1 to 5°C). The analysis is based on idealised simulations performed with an emission-driven Earth System Model, in which, after a ramp-up period (0.2K/decade warming), the anthropogenic carbon emissions are set to 0 for 300 years once the respective GWLs are reached. The authors investigate the evolution of the mean surface and subsurface temperatures (as well as global mean near surface air temperature) and relate the GWL-specific global mean temperature evolutions to balancing carbon and thermal effects. The GWL-dependent temperature evolutions are spatially resolved and linked to changes in the Atlantic Meridional Overturning Circulation (AMOC), which shows a GWL-dependent evolution after emission cessation. At low GWLs (≤4°C), AMOC strength has a tendency to recover during the stabilizations, while higher GWLs (>4°C) show continued AMOC weakening.
In addition to analysing changes in mean temperatures, the authors also document changes in sea surface temperature (SST) variability and its extremes. The authors find that with increasing GWLs, SST distributions also show increasingly widespread changes in variance. The authors continue to investigate the joint effect of warming and changing variance on marine heatwaves (MHWs), which they detect with a seasonally varying threshold and a GWL-specific fixed baseline approach. As such, they find that during the zero-emission runs, MHW frequencies increase at high latitudes, while decreasing at low latitudes.

Overall judgement:

The manuscript is clearly structured, well written and constitutes a valuable contribution to the growing literature investigating different aspects of the ZEC and its GWL-dependence. Furthermore, the results are presented in a concise way and the figures are of high quality. While I applaud the authors for synthesising the material into this concise manuscript, I see a few issues that I think need to be addressed or clarified before I can recommend the manuscript for publication.

Main comments:

1. Throughout the manuscript the spatially resolved SST evolution is not explicitly shown. Instead, Figure 4 shows a correlation with the AMOC index, which is difficult to interpret, given that the AMOC trend also changes sign across the stabilised GWLs (SWLs, Fig. 3). I would like to see maps for the SST anomalies (relative to the mean of the transient period at each GWL, i.e. the TWL), ideally for each of the early, mid and late stable periods. Otherwise, maps of the local SST trends could show the spatial patterns in SST tendencies during the stabilizations.
   I am furthermore not sure I understand why the correlation patterns in Figure 4 change so drastically (including in their sign) for different stabilised GWLs (ie., SWLs). This requires a more in depth discussion. If there is a physical link between local SSTs and AMOC, is there a reason why that link changes in sign at different SWLs (comparing for instance the sub-polar North Atlantic in 02Kpd-20 vs. 02Kpd-35)? Are the correlations significant? Are the correlations computed based on daily data or based on annual or longer averages, to reduce the influence of autocorrelation? Are they computed over the full 300 year stabilisation periods?
   That the correlation pattern looks different for the ramp-up appears sensible, as the SST includes the forced warming signal and not just the SST response to the AMOC reorganisation (and other adjustments) during the stabilisation period. That being said, I do not understand how a positive correlation can be achieved in the tropics during the ramp-up. As AMOC strength decreases during the ramp-up (see Figure 3), this implies to me that tropical SSTs are also cooling during the ramp-up which I don’t think is happening.

2. I am intrigued by the authors assessment of SST variance in addition to changes in the mean, which I think is a very useful analysis (Fig. 6, Fig. 7, Fig. S1-S5). However, I have some concerns regarding the employed methodology. Foremost, I require some clarification regarding the assessment of changes in the variance. Based on the methods and results section, I am under the impression that the assessed SST distributions are based on the time series from which solely the climatological mean seasonal cycle of the respective reference period is removed. Yet, to filter out changes in variance one also needs to remove long-term changes in the mean. Otherwise, these long-term changes of mean SST project onto the width of the SST distributions, which are then interpreted as changes in variance. This is particularly relevant for the ramp-up with its marked warming trend, but also the 300yr long stabilisation periods can be subject to low frequency variability.
   A separation of the SST signal into a low-frequency, an annual and an intra-seasonal component might be helpful.

3. The Kolmogorov-Smirnov test, which is used to check whether the transient (TWL) and associated stabilised (SWL) time series show different SST anomaly distributions, requires independent samples. However, SST anomaly time series generally show high autocorrelation (up to timescales of weeks to months), which I expect in this case to lead to too frequent rejections of the null hypothesis (Lanzante 2021). I have the same concern for the Welch t-test (testing the differences in the mean) and for the Levene t-test (test for the differences in the variance).
   If accounting for this autocorrelation leads to larger ocean areas for which the null hypotheses cannot be rejected, Figures 6, 7, S4 and S5 would benefit from the inclusion of additional colours or hatching indicating these non-significant changes.

4. In Section 3.2 the authors document that a large fraction of the ocean experiences increases in variance of SST anomalies between the TWL and SWLs, which I find striking. How large are these variance increases?
   If the statistical significance of the variance increases holds, given the above raised concerns regarding the variance computations and significance testing (main comments 2 and 3), it would be beneficial to discuss these results. What are potential drivers of the increase in the variance? Are surface mixed layers much shallower during the SWLs relative to the TWLs, so that they are more reactive to surface forcing? Are there other potential reasons? Or does the increase in variance rather stem from the low-frequency variability, which has more time to emerge in the 300yr stabilisations compared to the 21yr ramp-ups (see main comment 2)?

5. The assessment of MHWs relies on the (seasonally varying) 90th percentile threshold, which is a commonly used methodology (Hobday et al., 2016). Also the choice of a fixed baseline, using the transient warming level (TWL) period from the ramp-up as reference is justifiable. However, my concern with this approach is the following: Since the ramp-up experiences a warming trend of 0.2K per decade, the near-surface air temperature warms by ~0.4K over the 21yr-TWL. This warming trend, which also concerns SSTs, will bias the 90th percentile threshold to the later years of the reference TWL period. Since the GWL simulations are branched from the centre of the TWLs, the extreme threshold is therefore likely biased high, leading to a systematic underestimation of MHWs in the GWL simulations.
   Local internal variability, as well as the shorter 21yr periods used in this study compared to the typically used 31yr baselines in MHW studies (Oliver et al. 2021), likely alleviate this bias. Nevertheless, it would be insightful to discuss this point.

6. The authors show in Section 3.2 that not only the SST means but also the SST variances change in the zero-emission simulations (under the assumption that main comments 2 and 3 regarding the variance calculations and significance testing do not fundamentally change the result). Detecting changes in MHWs in Section 3.3 using the fixed baseline mixes the influences of these two factors (changing mean and changing variance), which the authors rightfully point out. A logical extension could be to use a “shifted baseline” in which the MHWs in the early-, mid- and late-stable periods are detected based on baselines corresponding to their respective periods. The difference between fixed and shifted baselines could then help to attribute changes in MHWs to changes in the SST mean and the SST variance, which would nicely complement the preceding analyses of SST anomaly

distribution changes.
Furthermore, a discussion of the underlying mechanisms explaining the shifts in MHW statistics would be beneficial. For instance, are the increases in the Southern Ocean MHW frequency in the mid- and late-stable periods (Fig. 8) linked to the continued AMOC decline? If so, why does it also show up in the 1.5°C and 2°C stabilisations, in which the AMOC recovers?

Other comments:

1. Figures 1, 2, 3, 5: It could be helpful for readability to colour the ramp-up lines in a different colour than red. Also the markers indicating the branching could take on the colour of the individual SWL line. This would facilitate readability, as, interestingly, the SWL time series partially follow the ramp-up trajectory.
2. L121: "taken as the mean of the esm-piControl simulation" - is this referring to the full piControl, or to the 31yr mean centred around the year of the start of the ramp-up?
3. L124: "… Is started on January 1st of that year" - I am under the impression that it is the following year if this follows the protocol from Jones et al. 2025 (L186-187)
4. L139: A brief discussion of why 21-year periods are used, even though MHW studies usually use 31-year baselines, would be beneficial.
5. L160: "(ΔOHU) was globally integrated" - derived from the sea surface heat fluxes?
6. L176: "an increase of each of these metrics would imply a greater risk of facing extreme events" - in the current formulation this statement depends on the definition of the extreme (e.g., fixed vs. shifting baseline) and "likelihood" might be a better word than "risk" in this context.
7. L196-L198: Is the percentile calculated using the 11-day-of-year rolling window before the smoothing with the 31-day-of-year running mean, as in the definition of Hobday et al., 2016?
8. L211: "… and 300 years for SWL300)" - I don't seem to find any results reporting MHW metrics over the full SWL periods.
9. Figure 2d-f: the piControl reference could be interesting to include for comparison
10. L241: "caveats from centennial cooling…" - is this a caveat because this cooling happens also in the piControl? (although the shown time series is not extending so far)
Why would these simulations need to be viewed differently than the others?
11. L245: It seems that the imbalance in Figure 2f is showing not only for 02Kpd-45 and 02Kpd-50, but also for 02Kpd-40. Yet, this simulation does not seem to show a cooling trend (Fig. 2c, Table 2). How can this be reconciled with the argument brought forward here, that the imbalance is the cause of the cooling?
12. L265: "… circulation-warming space (8Sv, 4.5K, see Fig. 3b)" - Figure 3b seems to have been dropped.
13. L275: "exponential fit equilibrium timescale t90 is 38 years for the subsurface layer" - Does this refer to the shallow subsurface depth range (10-70m) shown in Figure 5a? The evolution looks rather linear (with some variability), and the unexpected(?) decrease in timescale for the 10-70m range when going from +2.0K to +5.0K (38 vs. 21 years), makes me wonder whether fitting an exponential is the best approach.
For the deeper layer (100-500m) it seems to be reasonable.
A discussion of these results would also be beneficial and the fitting of exponential functions could be added to the methods section for clarity.
14. L349: "Pearson correlations between zonal early stable seasonal cycles and those from fifteen successive 21-year periods…" - it is not really clear for me what is correlated here. If it is a zonal mean (as indicated in the caption of Figure 10, would this not require a line of correlations for each latitude? Or are the correlations then averaged over all latitudes?

What is meant by “spatial Pearson correlation” in Figure 10 if seasonal cycles are correlated?

15. L402: “This finding highlights that the standard 20-year period …” - for MHW studies the typical baseline duration used is 31 years.

Grammar, typos and readability:

1. L8: “remain poorly understood”
2. L10: “arising” instead of “arisen”
3. L28-30: sentence difficult to understand
4. L53: “outweigh” or “dominate” instead of “overweight”
5. L54: “attributable to anthropogenic climate warming” - only in part. Extremes such as MHWs can also be driven by internal or natural variability.
6. L71: “and contrast between the transient …”
7. L108: “following the CMIP6 protocol”
16. Table 1: line “GWL +2.0°C” and column “Mid stable (SWL21)” should probably read “2089-2109” instead of “2089-2019”
17. L155: “We also evaluate in depth ocean temperature” - rephrase to “We also analyse subsurface ocean temperature”. I would also refer to the 10-70m and 100-500m depth layers as the “shallow” and “deep” subsurface layers, respectively.
18. L161: “positive fluxes when directed towards the atmosphere”
19. L163: “in-depth”, i.e. the “subsurface maximum” (below a particular depth level?)
20. L172: “first two statistical moments”
21. L192: “choice of baselines, thresholds as well as duration of [threshold] exceedence, …”
22. L199: “since the relevance…”
23. L227: “GMSAT shows a relatively good stabilisation” - maybe rephrase
24. L273: “due to quick GMSAT equilibration, subsurface ocean temperatures take longer …”
25. L276: “… while reaching centennial time scales for the deeper layer”
26. Figure 5: “Subsurface temperature evolution…”
27. L328: “… mid- and late-stable periods display more abrupt changes” - I don’t think abrupt is the right work here, as the system also had more time to evolve.
28. L363-L365: “This implies distinct mechanisms: …” - This sentence is not clear to me.

References:

Hobday AJ, Alexander LV, Perkins SE, Smale DA, Straub SC, et al. 2016. A hierarchical approach to defining marine heatwaves. Prog. Oceanogr. 141:227–38

Jones, C., et al.: The TIPMIP Earth system model experiment protocol: phase 1, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-3604, 2025.

Lanzante, J. R. (2021). Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnov and Kuiper's tests. International Journal of Climatology, 41(14), 6314–6323. https://doi.org/10.1002/joc.7196

Oliver ECJ, Benthuysen JA, Darmaraki S, Donat MG, Hobday AJ, et al. 2021. Marine heatwaves. Annu. Rev. Mar. Sci. 13:313–42