Response to reviewers

Review CC2

Thank you for your evaluation of our manuscript and the detailed questions on the model strategy and simulated processes. Please find a detailed response to your comments below.

1 Modelling

Forcing conditions are applied to the model and presumably include heat, currents and air circulation, which are all mentioned in the study without explicitly reporting the full scope of said conditions.

Simulations are said to be performed using the VIKING20X following the OMIP-II protocol, prescribing six consecutive simulations spanning the 1958 to 2019 time-frame, with the first one initialising from WOA13 data and oceanic conditions at rest, while each of the following cycles are initialised using the final oceanic conditions of the previous cycle; each cycle is then extended up to 2023 to analyse the 1980 to 2022 time-frame. In particular, only the first and sixth cycle of each series are analysed: I infer this is done to observe an immediate response to the forcing conditions in the first cycle and the influence of model drift and model spin-up in the last cycle. While I believe this is needed due to the limited timeframe in which data is available I could not find evidence of the validity of this cycle-based approach for the simulations either in this paper or in the provided reference (Tsujino et al., 2020), as after a new cycle has started oceanic conditions at a certain time t in the time-frame provided (1958-2019) would instead be mapped in the simulation to a time $t' = t + (n - 1)\Delta$, where n is the number of the cycle being computed and Δ is the length of the time-frame, to my understanding.

In general, all cycles allow us to infer the response to the surface forcing, as they are the same in all cycles. The surface forcing itself is a state-of-the-art dataset used by numerous modeling groups to study the past evolution of the ocean, which is described in *Tsujino et al. (2018)*.
The first cycle is here used to decipher the role of model drift. Although this is a simplification, the temporal evolution of the first cycle contains forced variability (related to the surface forcing), intrinsic variability (related to stochastic processes) and an adjustment to the initial conditions (model drift). Cycling through the atmospheric forcing multiple times allows for the model to dynamically adjust, reducing the drift after initialization. Each cycle has different initial conditions that are closer to the model's response to the applied surface forcing. Therefore, in an ideal case, the temperature evolution of the 6th cycle would only reflect forced and intrinsic variability. In reality the deep ocean may have not reached this equilibrium yet. Nevertheless, in the 6th cycle the impact of model drift is clearly reduced compared to the 1st cycle. Of course there are other spin-up strategies, but the goal is always to get closer to the model's equilibrium state, such that any spurious trends are minimized. The strategy we follow here is suggested by the OMIP-II protocol and indeed described and justified in Tsujino et al. 2020. The need for multiple cycles is for example explained by Tsujino et al. 2020: "However, in preliminary JRA55-do-forced (OMIP-2) runs conducted by many modeling groups, decline and recovery of the Atlantic meridional

overturning circulation (AMOC) occurred during the first few cycles before it reached a quasi-steady state.“

2 Results

The influence of geothermal activity on MHWs is not mentioned in the paper, and as such I would like to inquire if it is speculated to be noticeable, especially on bottom MHWs, or would stable geothermal activity not impact MHW formation due to their statistical definition?

This is an interesting question. Stable geothermal activity would not lead to a MHW as you point out. Locally the temperature would be warmer than in the surrounding, but constant in time. Nevertheless geothermal activity is very unlikely to be constant and could be related to MHWs in very active regions, e.g. around the Mid-Atlantic Ridge. However, it would still be a very local process that is way too small to be simulated in the model with a resolution of 3-5 kms. Additionally, the temporal variability of this heat flux in specific locations is not known and therefore it is not possible to test its impact in our model configuration.

The effects of model drift are shown to be greatly reduced in the linearly-increasing baseline. Being model drift defined as the adjustment of the simulated environment to unknown initial conditions, could it be argued that the primary effect of these unknown initial conditions is the temperature rise, thus reducing the model spin-up time needed, and that other lesser effects of said conditions are higher-order corrections?

If we understand you correctly, the question is whether a linear baseline approach reduces the spin-up time. Indeed the main effect of the initial conditions is to introduce a long-term adjustment that is mostly linear. There are also some non-linear effects, as evident in figure 3 for example. Nevertheless, these effects are smaller and can often be neglected. Thus applying the linear baseline in a model experiment with a shorter spin-up is possible (at least in our model) without introducing major errors (when the 6th cycle is regarded as our best estimate).

The simulations performed with this new protocol are said to not be decidedly more realistic than previous ones. Since the main difference from previous models is the impact of mesoscale dynamics, which could either have a cumulative impact or be averaged out over larger portions of the ocean, could these lower scale dynamics be seen as higher-order terms in the model approximations? If so a convergence interval should be defined, where the model can be argued to be more realistic, while outside of it higher-order corrections may not yield better approximations.

The model experiments are more realistic than previous ones used in many aspects. In particular this is related to the resolution of the model, which for example allows for a much more realistic path of the Gulf Stream and North Atlantic Current compared to coarser resolution models. This is one aspect that does not average out and can not be adequately represented with an eddy parameterization in a coarser resolution model. A correct GS/NAC path is very important locally for the characteristics of MHWs. Additionally, individual mesoscale eddies can generate MHWs (see for example Großelindemann et al. 2022). Such MHWs would be missing if an eddy parameterization is used in a coarser model, which can introduce the net effect of eddies on tracer gradients, but not the effect of individual eddies.

To define a proper convergence interval where the MHW statistics are no longer sensitive to the grid resolution, one would need the same experiment at various resolutions. However, we only have two resolutions (0.25° and 0.05°) available. In any case, when looking at extremes, non-linear processes are likely important and as shown in our study there are major differences between the two resolutions. This clearly indicates that coarser resolution models do not capture the effects of mesoscale variability on MHWs.

Being MHWs defined as events lasting at least five days, and since MHWs divided by less than two days are considered the same MHW, would this merging of MHWs cause problems for the MHW frequency data in areas where they tend to have longer durations along the length of the simulations?

As you point out, the duration and frequency are often directly related in particular as MHWs become very long. A MHW that lasts 365 days can, by definition, only occur once a year. Allowing for a gap of 2 days therefore leads to longer MHWs, but less MHWs. This however, is not necessarily a problem. An ecosystem may not recover from high thermal stress within 2 days, thus it is more reasonable to merge events. Using 2 days instead of any other gap is of course a choice and will often not be exactly the time an ecosystem needs to recover (and to allow the assumption that two events are independent). As our aim here is to provide an Atlantic wide three dimensional dataset as a complement to already available datasets for the surface, we choose to apply the common definition allowing for a 2 day gap.

The heat budget present in this study doesn't contain, at least from what is shown, the influence of the night-day cycle. Since MHWs at shallow depths are shown to be highly responsive to external conditions it could be an interesting forcing condition, but it may very well be averaged out over the multi-decadal time-scales used for the simulations, and it would outgrow the focus of the study on greater depth MHWs.

The day-night cycle is included in the model forcing (for example in the shortwave radiation flux). However, following the well-established definition of Hobday et al. (2016) the analysis is performed on daily means of the temperature. This is to maintain comparability to numerous other publications using the same approach. The most important motivation for using daily data is that temperature extremes are considered to have more severe impacts if they are sustained for sufficiently long time. Variations in the day-night cycle might be relevant for some species as well, but major ecosystem damages are expected when temperatures are higher than usual for at least several days.