In this study the authors used large ensemble simulations from one climate model to test to what extent pattern filtering approaches help to reduce internal variability in the dynamic sea level. They then discussed the benefits of using such approach to reduce uncertainties in pattern scaling of dynamic sea level change. This is an important research topic as large ensemble simulations are computationally expensive and usually we need to deal with limited or even single ensemble from climate model.

My main comment is that the reduced regression errors (residuals) in pattern scaling after applying the pattern filtering approach are well expected as the internal variability is reduced. I agree quantifying them is useful but the current manuscript fails to demonstrate more value for using such approach prior to pattern scaling, as claimed in the title and main message. Specifically, to what extent the application of pattern filtering could change the slope α of pattern scaling? Is there a significant change? Could you please show this change not only for global maps but also for time series in key regions as examples? Afterall this is what we really obtain and need from pattern scaling.

We thank the referee for their constructive comments. We agree that the comparison between the slopes α from raw and filtered simulations could further highlight the benefit of using pattern filtering approaches. We see substantial slope differences in places subject to non-linear mesoscale processes, such as strong western boundary currents (Fig. 1) (e.g., Gulf Stream and Kuroshio current; US east coast and Japan east coast, respectively). The maximum slope difference decreases with radiative forcing. Since lower radiative forcing means lower signal/noise ratio, noise (internal variability) can drive large differences in slopes between filtered and unfiltered results. On the contrary, a higher emission scenario is characterized by a higher signal/noise ratio, as noise exerts a less important control on slope differences.

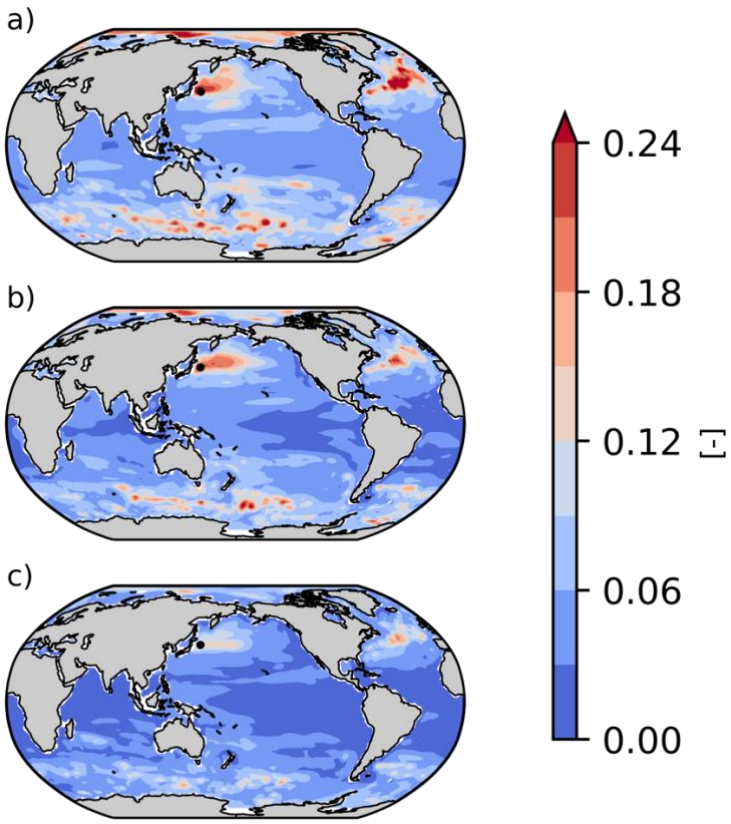

Figure 1. Maximum slope difference between unfiltered and LFCA-filtered realizations, considering all 100 MPI-GE members, for RCP26, RCP45, and RCP85 (a, b, and c, respectively). Black dot to the east of Japan represents location for figures 2, 3, and 4.

We will include a plot (similar to Fig. 1) showing differences in slopes for distinct RCPs considered here. Also, as suggested by the reviewer, we will also include (in supplementary) key regions where changes in slope are substantial. As an example, we are including here a point in the Kuroshio current (east coast of Japan, see black dot in Fig. 1)., comparing an unfiltered, a 30-yr moving mean, and a filtered realization for RCP 2.6, 4.5 and 8.5 (Figs. 2, 3, and 4; respectively). Fig. 2 shows that while unfiltered and 30-year moving means are quite similar, the filtered case shows a positive and much steeper slope (for a single realization, as an example). These differences are caused by internally generated variability and the removal of data points to compute the 30-year mean to remove part of the temporal variability. Slope differences get smaller as radiative forcing increases (Fig. 2 vs 3 vs 4), as also shown in Fig. 1

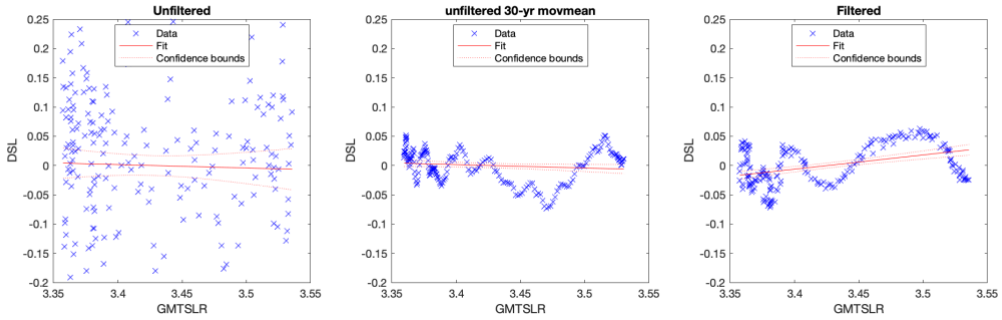

Fig. 2. Linear regression model of dynamic sea level (DSL) and GMTSLR for an unfiltered, a 30-year moving mean, and an LFCA-filtered RCP 2.6 realization (left, middle, and right panel, respectively).

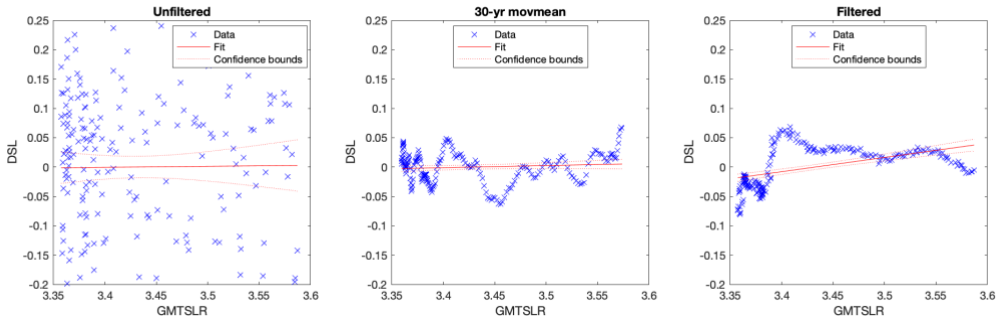

Fig. 3. Linear regression model of dynamic sea level and GMTSLR for an unfiltered, a 30-year moving mean, and an LFCA-filtered RCP 4.5 realization (left, middle, and right panel, respectively).

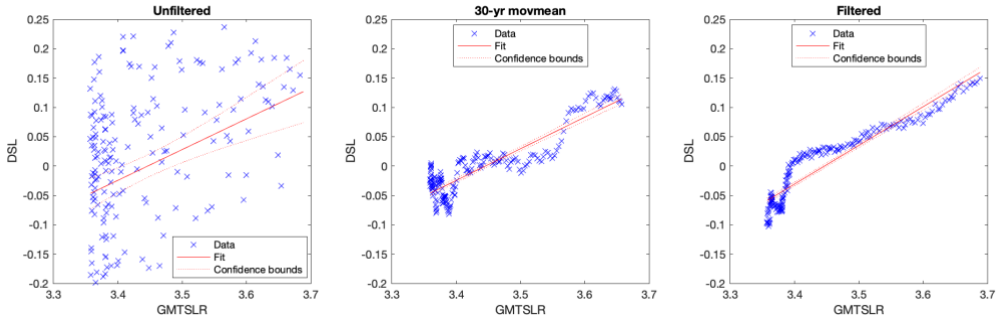

Fig. 4. Linear regression model of dynamic sea level and GMTSLR for an unfiltered, a 30-year moving mean, and an LFCA-filtered RCP 8.5 realization (left, middle, and right panel, respectively).

Some minor comments below,

L21 “model disagreement” is not straightforward here – please consider rephrasing. In the context of last sentence does it refer to “climate model” or “statistical model”? should “disagreement” be “uncertainty” here?

We refer to disagreement between statistically modelled ocean dynamic sea-level change and simulations coming from the respective GMC. To avoid confusion, 'model disagreement' will be changed to 'statistical model error'.

L26 "MPI-GE" might not be familiar to some readers

MPI-GE will be introduced as "Max Planck Institute Grand Ensemble (MPI-GE)" both in the abstract and main text in the revised manuscript.

L26 "so that internal variability is optimally characterized while avoiding model biases" – please consider rephrasing. We can never avoid the model bias issue. My understanding is when using single model large ensemble simulations, the externally forced signal is optimally characterized, which provides important basis to test pattern filtering methods.

The reviewer is right about model biases: it will still be a problem even when using a single model. What we were trying to emphasize here is the benefit of using single-model large ensembles instead of utilizing same-forcing simulations from different models. The former allows us to optimally characterize the externally forced response within a model, whereas the latter could include model biases as externally forced response. We will rephrase this sentence to emphasize large ensemble simulations allows to optimally characterize the externally forced signal within a model and forcing scenario, instead of saying that using them allows us to avoid model biases.

L27 "pattern filtering" do you mean the "two pattern recognition methods (L23)" or specifically the "signal-to-noise maximizing EOF pattern filtering (L24)".

We refer to both methods. We will clarify this in the text.

L66 "natural" should be "internal climate" as used in most other places – please check throughout the manuscript for this.

We agree with the reviewer that, as written in the original paper, natural and internal climate variability seem interchangeable when they are not. We will check these terms throughout the paper and make modifications accordingly. We will also include a definition of internal climate variability as suggested by the other reviewer, reducing ambiguity. In addition, we noticed we refer to internal climate variability many times, so we will introduce the abbreviation ICV to increase readability.

Figure 3 It's unclear (1) how the number of ensembles needed is calculated; (2) what does "forced response variance" refer to. Could you please make connections to equations in section 3?

First, we would like to clarify that we calculated the required number of ensemble members (realizations), a not the number of ensembles needed. (1) The number of ensemble members needed to explain a certain level of variance of the forced response is based on the coefficient of determination r2 between the two datasets considered. Here, chose the 80% of the variance following similar studies, but other arbitrary level could be chosen. The procedure we to took is as follows:

i. Create two subsets of the 100-member ensemble, with 50 members each.
ii. The forced response is estimate from one of the 50-member ensembles using all members in the subset. We used this forced response as reference.
iii. The forced response is also calculated from the other 50-member subset but instead using all 50 members as in (ii) the number of members is increased from 2 to 50 in an iterative process. We call this subset the testing subset, as it is the one used to estimate the number of ensembles needed to explain a certain level of variance in the reference (step ii) subset.
iv. The number of required members is computed as follows. We start with only 2 members, which are used to estimate the forced response in the reference subset. We compare both forced responses (2-member testing subset vs 50-member reference subset) by means of the coefficient of determination (r2) which tells us about the proportion of variance that is shared between the two subsets. We do this on a grid-point basis and see where the 80% level is exceeded. For those grid points where the threshold is not exceeded, we do the same comparison but adding an additional member to the testing subset (i.e., 3 members). Again, we check where the 80% threshold is exceeded when an extra member is considered. We continue this procedure by adding more members until we reach 50 members (the maximum in the testing dataset).
v. We do this comparison when either the forced response is calculated by averaging (Fig. 3a in manuscript) or S/N M EOF pattern filtering (Fig. 3b in manuscript).
vi. To avoid sampling bias, we repeat this analysis several times by randomizing the initial 100-member ensemble.

We hope it is clearer now. A couple of final notes for this answer.

- When we say forced response variance, we refer to the proportion of the variance that is shared between the two subsets being compared here. We'll clarify this in the text.
- It is difficult to make connections with section 3 (Methodology), since none of the equations used there has relation to the calculations performed here to estimate the require number of ensembles members to explain the forced response variance within a subset.

Although we have attempted to explain how the calculation was performed in the text (pag. 333 to 342), we will expand such explanation so that the interpretation of the results in Figure 3 is easier.