Reply to the reviewer #1

This study introduces an impressive new dataset that is available for community use with the MIROC6 model. This dataset consists of large ensembles of historical and future projections with multiple scenarios as well as single forcing simulations. The primary purpose of the paper is to introduce these simulations but it also presents some cursory, but useful, analyses of changes in global temperature, precipitation and their extremes and some assessment of non-linearity in the single forcing simulations and analysis of the number of members required to detect changes or differences between scenarios. Overall, I think this is a useful and well written study that introduces this important new dataset and I have only minor recommendations to consider before publication.

Thank you very much for your useful comments.

l83: It could be worth providing a bit more information about the piControl and the initialization dates. Firstly, it could be worth stating whether the piControl is still drifting at this stage and some more specifics about the initialization dates e.g., were they spaced by a certain number of years?

The piControl run is stable (Fig. 3 of Tatebe et al. 2019). We will involve the list of initial conditions as a table.

L87: It can sometimes be a bit confusing where biomass burning aerosols are represented. I'm assuming that they are included in the anthropogenic aerosol contribution? Even though there is a natural component to that. It might be worth being clear about this.

Although biomass burning aerosol emissions include both anthropogenic and natural components, anomalies from piControl (involving the 1850 emission of biomass burning aerosols) can be used to estimate the anthropogenic component.

l120-126 and Fig 3: I'm not sure what the motivation is for doing this assessment of non-linearity by using only samplings of 1 member. It may be that there is a true non-linearity but you can only see it in the ensemble means. You could do the same analysis but sample N members with replacement from each ensemble, where N is your original ensemble size, to determine whether there are any non-linearities that can be detected with the ensemble means.

Thank you for the useful advice. For the revised version, we have computed the max-min ranges of 1000 ensemble average values of randomly sampled ensemble members with replacement. The range of the ensemble mean of historical runs overlap with the range of the sum of the ensemble average of individual forcing experiments. Therefore we have not found any statistical evidence of non-linearities in the ensemble mean values.

l137: At the introduction to Fig 5, it might help readers to remind them what time period is being considered. I think it's 2000-2020 minus 1850-1900?

We will denote “2000-2020 minus 1850-1900” in the caption of Fig. 5.

l147-151: I think this text is describing the behavior of the hist-nat+ssp245-nat run in Figure 6, but it's not entirely clear. Maybe reference that part of the figure when referring to the solar and volcanic contributions.

This text is describing the solar and volcanic forcing in the future simulations (2015-). We will made it clear in the text as “For future simulations from 2015, volcanic …” and will draw only the period of 2015-2100 in Fig. 6.

l210: It seems like another possibility beyond the interannual external forcings is inacuracies in the use of a linear trend? If so, that could be mentioned too.

Thank you for the question. First, we would like to clarify the sentence, “… the forced response of the Niño-3.4 SST is not sensitive to interannual external forcing, such as volcanoes”. As seen in Figures 1 and 2, the ensemble averages (forced responses) of the global mean annual temperature and precipitation are substantially affected by natural external forcing such as volcanic eruptions. Thus, the “SMTR estimate” method (the linear trend is removed in each ensemble member) is not appropriate for estimating their internal variability amplitudes (gray and green plots in Figures 11a and 12a). This issue has already been mentioned in the other part. In contrast, the internal variability amplitude of the Niño-3.4 SST is close between the SMTR and MMMR estimates when using the 50 ensemble members (Figure 10a), despite that the natural external forcing such as volcanoes could potentially induce an interannual-scale forced response in the Niño-3.4 SST. This implies that such forced response is relatively small and thus the SMTR estimate could derive an accurate amplitude of the Niño-3.4 SST internal variability.

As for potential inaccuracies in the use of a linear trend such as noise influences, we think that these have been already implied as follows: "the second method, which requires a large ensemble, is more appropriate for determining the internal variability component since it is not contaminated by residuals from detrending methods."

l224: I got confused by the wording here.   You refer to a "single-member estimate" but then proceed to discuss the method, which doesn't sound like a single member estimate at all.   The description sounds like an "N member estimate".   Suggest clarification.

We will change the wording to describe the two methods. After suggestion by reviewer #2, we will use "single-member-trend removed (SMTR) estimate" and "multi-member-mean removed (MMMR) estimate" through the manuscript.

l246: Presumably some measure has been chosen to quantify whether it has been "degraded".   Suggest being clear by what measure you are using here.

We will remove the specific threshold number in the corresponding text. Instead, the text will be rewritten in a general way as follows: "a smaller ensemble size (e.g., N=10) results in a higher probability of the underestimated standard deviation by the MMMR estimate, demonstrating the necessity of a large ensemble to evaluate the internal variability amplitude."

l310: Again, it seems you need to have chosen some threshold to quantify whether the amplitude of the internal variability is underestimated.   Suggest being clear about how you have determined that.

Similar to the above response, we will remove the specific number and rewrite the text in a general way as follows: “a large ensemble is necessary to avoid underestimating the amplitude of internal variability relative to the ensemble mean.”

L315: There is an accompanying single forcing large ensemble for the CESM2-LE which I think would increase the number of years of simulation for CESM2 (https://www.cesm.ucar.edu/working-groups/climate/simulations/cesm2-single-forcing-le)

Thank you. We will denote the sum of the following simulations (32675 years):

- historical + SSP370: 100 members of 1850-2100 = 25100 years
- GHG only: 15 members of 1850-2050 = 1515 years
- Anthropogenic aerosol only: 20 members of 1850-2050 = 2020 years
- Biomass burning only: 15 members of 1850-2050 = 1515 years
- All but GHG, AAER and BMM: 15 members of 1850-2050 = 1515 years
- All but AAER: 10 members of 1850-2050 = 1010 years

Typo's/wording:

l58: suggest changing "and" between the reference to the biomass burning simulations and the greenhouse gas simulations to "or" since it is not both that are time evolving.

It will be changed.

l131: "TX" --> "Tx"

We will correct it.

l157: Here, and throughout, there's some inconsistency as to whether you refer to "ssp" or "SSP" and "ssp245" or "SSP-2.45". Suggest being consistent.

In CMIP6, SSP2-4.5 indicates the concentration scenario, and ssp245 means experiments under the SSP2-4.5 scenario. For example, please see Gillett et al. (2016, https://doi.org/10.5194/gmd-

9-3685-2016). Although it may be slightly confusing, we will follow the usage of those terms in CMIP6.

l252: "variabilities than the best" --> "variabilities compared to the best"

We will correct it.

l319: "federation grid" --> "grid federation'

We will correct it.