We would like to thank the reviewer for providing helpful comments on this manuscript. The suggestions will help improve the manuscript in a revised version. We provide a response to specific comments below.

The manuscript by Haywood et al. presents results from the Pliocene Model Intercomparison Project Phase 2 and follows on an earlier Pliocene model intercomparison published in 2013. Despite the title, the manuscript focuses on two issues: Pliocene large-scale climate features and earth system/climate sensitivity. The study has the potential to be an important contribution. As presented, the study has several shortcomings, described below. My recommendation is to revise the manuscript, eliminating the analysis of earth system sensitivity and expanding the analysis of large-scale features and the comparison with proxy data.

The manuscript is intended as an introduction to the main results from PlioMIP2 and includes a large range of features in temperature and precipitation across different timescales and spatial scales. We also present a comparison with proxy data and consider the relationship between climate sensitivity and the Pliocene climate anomaly.

We agree that it would be possible to move the analysis of Earth System Sensitivity and Climate sensitivity to another paper. However, we do not agree that this is necessary or indeed desirable. The inclusion of this analysis provides a useful link between PlioMIP2 output, Pliocene proxy data and future climate change, which broadens the relevance of the manuscript. Removing this analysis would be to the detriment of the manuscript.

However, we acknowledge that the ESS/ECS section can be improved. We will simplify this section in the revised version of the paper. In the previous version complications arose from

1. *The EC-Earth3.1 model, which had ESS < ECS – which was unreasonable. This model has been now withdrawn from PlioMIP2 by the EC-Earth modelling group due to recently determined problems in sea ice sensitivity.*
2. *The CCSM4 models not being internally consistent. We had three very different representations of the Pliocene climate from CCSM4 due to different groups using different model parameterisations and different ways in which the Pliocene boundary conditions were implemented. We therefore took decision to average these models together to investigate the relationship between ECS (which was the same in all the CCSM4 models) and ESS (which was different in the CCSM4 models).*

In the revised version of the paper we will include all models in our ECS/ESS analysis, including treating all versions of CCSM4 as separate and distinct models. This will make the ECS/ESS analysis consistent with the rest of the paper, where all versions of CCSM4 were treated as separate models.

In addition, since the previous version of the paper we have received two further contributions to PlioMIP2. These are EC-Earth3.3 and CESM2. In the revised version we will include these two additional models, and this strengthen the results seen in the initial submission. In particular, the inclusion of these additional models strengthens the link between ECS and ESS and makes this section more robust.

We agree with the reviewer that the title did not accurately reflect the contents of the paper. In the revised version we will change the title to show that the paper deals with both large-scale features of the Pliocene climate and climate sensitivity.

General comments

Discussion paper

As an analysis of a model intercomparison project, the manuscript is fine. It reports the ensemble average and range across a spectrum of large-scale features including global, zonal and seasonal temperature; polar amplification; SST gradients, and precipitation rate, and compares them to pre-industrial conditions. These results are interesting and relevant, but the analysis is rather cursory. It would have also been interesting for the authors to include some direct comparisons with PlioMIP1 in the Results (for example, include PlioMIP1 ensemble means in the figures and description of results). In the Discussion, there is some speculation about why the PlioMIP2 results differ from PlioMIP1, but it's just speculation. In addition, the analysis of PlioMIP2 models lacks investigation of why large-scale climate features differ among PlioMIP2 models. Both of these are major missed opportunities.

The manuscript represents the first results from the PlioMIP2 ensemble and therefore we included many comparisons/analysis that subsequent studies will expand on. We acknowledge that the reviewer thought the analysis was too brief, however, the manuscript is an overview for a broad audience and already runs to a number of pages, hence we would not like to add a large amount of additional detail.

In a number of places, the text compares results with PlioMIP1, however we note from the reviewer's comment that a more quantitative comparison would be desirable. In the revised version of the manuscript we will add the PlioMIP1 results to appropriate figures (fig 1a, fig 1c, fig3, fig4a, fig6).

We agree with the reviewer that it would be very interesting to understand why large-scale climate features differ among different models. However, this is a very difficult question to answer and it is beyond the scope of this study to do this thoroughly for all diagnostics. In general, the models that have the greatest Pliocene warming also have the greatest published climate sensitivity, and this will be emphasised more in the conclusions. Where all models have been presented individually (i.e. fig 1a, fig4a etc) we will reorder the models in terms of highest ECS to lowest so that the reader can more clearly see how the diagnostics relate to each other.

The estimate of earth system sensitivity (ESS) (equation 1) is not explained or justified. There is no a priori reason to think that the ESS will scale as the ratio of ln (560/280)/ln(400/280). This scaling would be appropriate if CO_2 were the only factor changing between simulations. That this scaling is inappropriate is illustrated by the differences in ESS for CCSM4 (Table 2, compare CCSM4-2deg and CCSM4-UofT). In these two simulations with the same model, the Eo400-E280 SAT differs by 0.9 C due to difference in treatment of Pliocene boundary conditions. (No surprise.) The ESS estimate (equation 1) grows the difference to 1.8 C. Why would the same model respond so differently to an increase in CO_2 of 160 ppm? If the authors believe this result is justified, they must demonstrate it by running these two CCSM4 Pliocene simulations with 560 ppm CO_2. Given the shortcoming in the ESS estimate, all discussion of earth system sensitivity should be removed from the manuscript.

The reviewer makes two points here:

1. *That equation 1 is not explained or justified*
2. *That the CCSM4-2deg and CCSM4-UoT provide a very different response to the mPWP boundary conditions, and hence ESS (calculated from these different version of the same model) is very different when it should not be.*

To respond to point 1:

the ESS is simply

*DeltaR * ESS = DeltaT*
*Where DeltaR is 5.35 * ln(400/280)*

As far as the determination of ESS is concerned we are not doing anything unusual. As elaborated in Chandan and Peltier 2018 (this issue), the usage of this equation is justifiable on ESS timescale if we consider the various ice-sheet and GIA related changes as long-timescale earth system responses arising from CO2 changes in the first place. This is illustrated schematically as Figure 11d in that paper. To the extent this argument is correct, i.e. to the extent that non GIA related orography changes are small, and which appears to be the case from Dowsett et al. (2016), the above formula should indeed give an estimate of ESS even though on the LHS the forcing is just that from CO2 and on the RHS the total temperature difference includes contributions from various other factors. A minor caveat is that the non-GIA related changes are most pronounced in the infilling of the straits in Northern Canada and the Hudson Bay. But this would only significantly affect local calculations and is not expected to affect the results from the above equation which is based on a global mean.

To respond to point 2:

The reviewer notes that "In these two simulations with the same model, the Eo400-E280 SAT differs by 0.9 °C due to difference in treatment of Pliocene boundary conditions. (No surprise.)" Actually, this difference of 0.9°C between two very similar models was a surprise to us, and we spent a great deal of time trying to understand the exact differences between how the boundary conditions had been implemented in CCSM4-UoT and CCSM4-2deg. This is discussed in lines 205-215 of the manuscript. The reviewer's suggestion to run these two CCSM4 simulations with 560ppm CO2 would provide clarity to this issue, however these simulations are very expensive to run and were not a requirement for a modelling group to contribute to PlioMIP2 (Haywood et al. 2016 – this issue).

While we agree that some uncertainties remain about the CCSM4 models, we do not agree that this translates into uncertainties of our ESS methodology. The disagreement between CCSM4 models is apparent throughout the manuscript – not just in the ESS section. Even with this disagreement the strong correlation between ESS and ECS is sufficient that the PlioMIP2 ensemble and proxy data can be used to help constrain ECS. This will become more apparent in the revised version due to improvements in the ESS/ECS analysis.

Specific comments:

Introduction.

The Introduction could be improved. It does a poor job of justifying the rationale for conducting PlioMIP2. A strong case could be made that the PlioMIP2 offers an opportunity to evaluate climate models that have been strongly tuned for the present day, and to showcase advances in modeling since PlioMIP1. Instead, the Introduction (paragraph 2 specifically) is a clumsily written laundry list of all the publications that resulted from PlioMIP1. What's the point?

The reviewer makes some good points here. In the revised version of the manuscript we will rewrite the introduction to incorporate the reviewer's suggestions.

L. 156. The minimum integration length was specified as 500 simulated years. For many models, this is not sufficient time to reach equilibrium. I appreciate that the authors report the spin up time and net TOA radiation in Supplementary Figures and suggest that they add this important information to the manuscript.

The manuscript has been written to appeal to a broad audience, only some of whom will be interested in the spin up time and the net TOA radiation. We think it is better to keep this information in the supplement, where it can be found by interested parties without distracting other readers.

L. 164. Here and elsewhere (e.g. L. 473) the manuscript mentions the release date of the model, and even makes statements like "the model sensitivity is more strongly related to parameterization choices and initial conditions than the release date of the model". As the authors are well aware, model performance is related to the accurate representation of the dynamics and physics and has nothing to do with the date of release. Please remove these confusing and unnecessary comments.

The purpose of this analysis was to determine if developments in model physics lead to altered responses to Pliocene boundary conditions. In particular, whether newer (more recently released) models might show a different sensitivity than older models. We will clarify our meaning in the text.

L. 237. "Lack of consistency in the seasonal signal of warming..." An analysis of the global average seasonality is not very useful (or at least ambiguous) here since the seasons are out-of-phase between hemispheres. Please show hemispheric averages, or just the NH average. Also, this is a place where additional analysis would be appreciated to understand the reason for the model differences.

This is a good point. In the revised version of the paper we will just show NH average. We will also remove the seasonal correlation for ESS in figure 7.

L. 339. "...suggests that there are some inconsistencies between the way in which ECS and ESS were obtained." See my comments above about the estimate of ESS used in the manuscript.

The EC-Earth3.1 model (about which this statement was written) has been withdrawn from the PlioMIP2 intercomparison due to problems in sea ice sensitivity. This sentence will not be required in an updated version of the manuscript.

L. 344. “each of these models provides a different but equally valid realization of ESS...” I don’t understand this statement. Please elaborate or delete.

This paragraph will be removed in the revised version of the manuscript. This is because we will be treating all the CCSM4 models as separate models in the ESS section (in the same way we have done in the rest of the manuscript).

Data/Model Comparison. This section (lines 389-400) focuses on comparing the ensemble mean to the proxy data. How consistent are the models? It would be valuable to add a figure showing model agreement, the number of models that were within the criteria for a good fit. In addition, it would be valuable to estimate and report the goodness of fit for each model, as well as the ensemble mean.

We agree that additional information about how individual models compare to the proxy data would be useful. In the revised version, a figure showing how each model compares to the proxy data will be added to the supplementary information.

L. 410. Please calculate and report the global mean DSAT/ DSAT estimate from proxy data for comparison. There are a number of ways to do this that have been reported in the literature.

This is an interesting suggestion, but the spatial distribution of Pliocene SST data is too limited, in our opinion, to make such a process reliable.

Section 5.1. This section is quite interesting. In line 460-464, it is stated that there are differences between the Pliocene and RCP predictions. Please elaborate on these differences, in the same way that the similarities have been described.

Here we were discussing changes in boundary conditions between the PlioMIP2 simulations and RCP simulations. We will make this clear.

L. 489. “Previous DMCs for the Pliocene...” I don’t understand this sentence. Please clarify.

This sentence will be rewritten in a revised version.

L. 545. “...suggest that SST data from the Pliocene tropics has the potential to constrain model estimates of ECS...” The discussion that follows (L. 550-577) about equilibrium climate sensitivity is not robust. ECS is calculated from a handful of local points from the same regions and is justified because it agrees with the ECS reported in IPCC, the exact value that this analysis should be testing.

This discussion arose from two points:

1. *The Pliocene temperature anomaly at many model gridpoints is correlated with the model's ECS.*
2. *Some of the gridpoints where this correlation occurs have proxy data – providing an independent estimate of the Pliocene climate.*

While models provide a range of estimates of climate sensitivity, combining points 1 and 2 allow us to estimate climate sensitivity from proxy data at certain locations.

We did not intend to justify our method based on it agreeing with IPCC, rather we intended to compare our results with IPCC estimates (and also with earlier PlioMIP1 studies) for completeness. We will rewrite the text around line 571-572 to make this clear.

The additional models that have been added to PlioMIP2 since the initial submission improve the relationship between Pliocene temperature anomaly and ECS. This means that the number of proxy data points that can be used to estimate ECS will increase in a revised version, and will cover a wider region.

L. 629. This is not a conclusion of this study, and therefore shouldn't be included in the Conclusions.

We agree with this point and this conclusion will be removed in a revised version