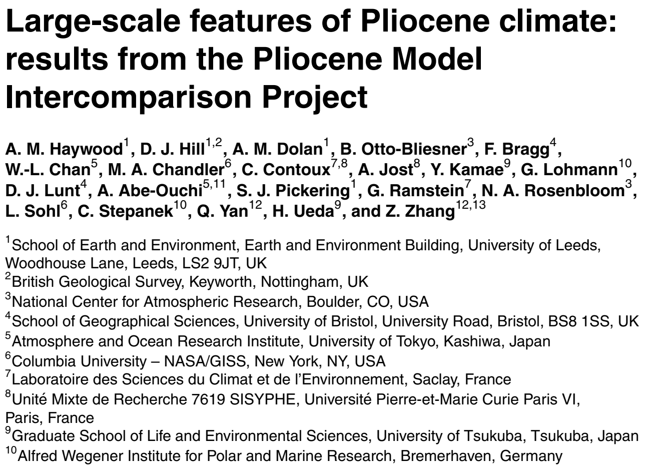
Review of Large-scale features of Pliocene climate: results from the Pliocene Model Intercomparison Project

reviewed by Matthew Huber



Large-scale features of Pliocene climate: results from the Pliocene Model Intercomparison Project

A. M. Haywood[1], D. J. Hill[1,2], A. M. Dolan[1], B. Otto-Bliesner[3], F. Bragg[4], W.-L. Chan[5], M. A. Chandler[6], C. Contoux[7,8], A. Jost[8], Y. Kamae[9], G. Lohmann[10], D. J. Lunt[4], A. Abe-Ouchi[5,11], S. J. Pickering[1], G. Ramstein[7], N. A. Rosenbloom[3], L. Sohl[6], C. Stepanek[10], Q. Yan[12], H. Ueda[9], and Z. Zhang[12,13]

[1]School of Earth and Environment, Earth and Environment Building, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK
[2]British Geological Survey, Keyworth, Nottingham, UK
[3]National Center for Atmospheric Research, Boulder, CO, USA
[4]School of Geographical Sciences, University of Bristol, University Road, Bristol, BS8 1SS, UK
[5]Atmosphere and Ocean Research Institute, University of Tokyo, Kashiwa, Japan
[6]Columbia University – NASA/GISS, New York, NY, USA
[7]Laboratoire des Sciences du Climat et de l'Environnement, Saclay, France
[8]Unité Mixte de Recherche 7619 SISYPHE, Université Pierre-et-Marie Curie Paris VI, Paris, France
[9]Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan
[10]Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany

Tuesday, October 9, 2012

Table of contents

Precís 1
Review 1
Summary of the main weaknesses 1
Specific issues and recommendations 2

Precís

Haywood et al describe the large-scale results of PlioMIP. This an interesting and important project and this paper does a good job of summarizing its main results. The study is impressive in its scope and the effort that went in to it. The use of both specified SST (Exp 1) and coupled models (Exp 2) is especially laudable. The manuscript covers a lot of ground, much of it well, but with some key gaps and some conceptual and statistical missteps that undermine the main conclusions. It is not publishable in its current form because of serious deficiencies in its handling of the model-data comparison.

I am suggesting moderate revisions that focus on : (1) including the presentation and discussion of some important variables not considered; (2) suggesting some alternative statistical treatments that may bring out the key areas of model-data agreement and mismatch better; and finally, as a result of (1) and (2) I suggest that (3) adjustment of the main conclusions and implications may be in order (but that's unknown at this point). I also make my case for specific discussions that would help a general readership better understand the methodology and it's strengths and weaknesses.

Review

Summary of the main weaknesses

- This study suffers from some sins of *omission*. Generally speaking, no information on the top-of-atmosphere (TOA) or surface residual energy budgets is presented. Such quantities, such as global mean TOA imbalance or the meridionally integrated zonal surface residual budget can provide key diagnostics that bear on issues such as equilibration, climate sensitivity, and implied ocean-atmosphere heat and water transports. This issue is relatively easily handled since these values are almost certainly ones that are saved in the PlioMIP archive and they are easily presented in tables or in line graphs. It is quite possible that this will affect the conclusions reached about the correlation (or lack thereof) of the PlioMIP sensitivity versus each model's "Charney" sensitivity, as well as the interpretation of evaporation, precipitation, and temperature gradient changes. I make some specific suggests about what to add and how in the Specifics Issues section. This will effect Figures 1, 2, and 3 .

- This study also is guilty of sins of *commission* in its handling of averaging and statistical analysis. The authors never come out and state what their assumptions are about the statistical distribution of the data and the model are, instead they seem to automatically equate a 2**σ** error with a 95% confidence interval without any discussion. This is just one symptom of a deeper problem. It seems they are implicitly assuming normal distributions for all their data, at least when convenient, and ignoring spatial co-variance, non-gaussianity, etc. It also appears that they are not adding error in quadrature when aggregating estimates, which is an odd choice for random error. In Figure 6, no discussion of what the limits of each regional box is, or how it was derived is presented. The difference between "Type 1" and "Type 2" error, is not considered. The authors have not plotting up the residuals and tried to understand the residual error structure. At a minimum the authors should state their assumptions about the underlying distributions and how that influences their statistical choice. More appropriately, they should actually test to see if their distributions are normal or not and act accordingly. This is not simply a matter statistical niceties. The authors make claims using this statistics that fly in the face of a simple perusal of the data in Figure 5, so a proper justification of why systematic model errors in Figure5 disappear after massaging in Figure 6. One useful diagnostic vis a vis the "Type 2" error issue would be to plot the model data anomaly in Figure 5, not versus latitude, but versus temperature (broken into N and S Hemisphere). This would almost certainly show a large systematic bias.

  - Probably the most troubling aspect of the handling of data in this paper is that it obscures the obvious bias in meridional temperature gradient that the models have (Figure 5). Various choices could be made about how to best present that discrepancy and it is not immediately obvious to me which is the best, but I can suggest one. Both separately and combined (for Exp 1 and Exp 2 results) take all the results presented in Figure 5 and aggregate them into bins: 30°N to 30°S, 30°S to 90°S, and 30°N to 90°N. Then present PDFs of those distributions, and calculate the usual statistics on them (mean, median, skewness, kurtosis). Use the histograms to generate the 95% confidence intervals (do not use parametric statistics unless you can show that the PDFs are normal). Then use that information to calculate the median tropical and extratropical (N and S hemisphere) biases and determine whether they are significantly non-zero. With that information in hand, then calculate the bias in the temperature gradient (N and S separately). This can be done with a couple simple histograms and a data table. My impression based on Figure 5 is that this effort will reveal a systematic bias in the meridional temperature gradient (it will be most clear when the terrestrial and ocean data are combined).

- It is always dangerous to guess what analyses will demonstrate, but for the sake of making it clear the conceptual basis for why I consider the diagnostics I suggested above are so important, I will make some guesses.

  - I think the authors may find that some of the Exp 1 simulations are not in TOA balance at 405 ppm CO_2. In that case, the simulations are (a) not in equilibrium, which bears mentioning, and (b) the model 'fails' one gross but fundamental test of its validity which is also important to mention.
  - With fixed SSTs (and an arbitrarily specified TOA balance) evaporation may be unrealistically enhanced or diminished (since it is the main way for the ocean surface to achieve balance when SSTs are fixed). When discussing the differences between the Exp 1 and Exp 2 results this should be highlighted and global mean E and P results should be shown, as well as latent heat transport in both Exp 1and Exp 2 configurations.
  - Some of the Exp 2 simulations may also not be in equilibrium at 405 ppm and this may affect the inferred climate sensitivity in Figure 1. At the least, the global mean TOA residual should be used to adjust the 'equilibrated' temperature value (utilizing the known 'Charney' sensitivity numbers), and a new Pliocene sensitivity calculated with that adjusted value.
  - The Exp 1 results will imply a set of ocean and atmospheric heat and water vapor transports that may not be reproduced in the Exp 2 studies. Specifically, the implied ocean heat transport in the Atlantic from the Exp 1 simulations is likely to be far apart from the actual heat transport prognosed in the Atlantic in the Exp 2 simulations. This is another important test of the models, if they consistently underestimate Northward heat transport in the Atlantic, this bears mentioning.
  - Part-and-parcel of what I'm guessing will be a failure of the coupled models to produce enough ocean/atmosphere heat transport is the inability of the models to accurately capture the weak temperature gradients in the proxies. I imagine that this may be a important conclusion/implication to mention (if it is true). This also suggests that—arguments in the paper to the contrary—a refinement of the 'time slice' of the model-data comparison interval may not significantly improve agreement because a similar failure of models to produce weaker temperature gradients in warmer worlds is found from universally from the Cretaceous through the Miocene. Again, this might alter the conclusions and implications of this paper if it is true.

Specific issues and recommendations

- **Abstract**: However, data/model comparison highlights the potential for models to underestimate polar amplification. To assert this conclusion with greater confidence, limitations in the time-averaged proxy data currently available must be addressed.

Actually, I think there is very clear evidence that the models are not matching the data and the failures are straightforward to present. After completing the analyses suggested here that the abstract may need some work.

- **2975:** In both Pliocene experiments the atmospheric concentration of carbon dioxide (CO2) was set at 405 ppmv. This value falls within the uncertainty limits of current CO2 proxy records (e.g. Pagani et al., 2010; Seki et al., 2010; Bartoli et al., 2011). All other trace gasses were specified at a pre-industrial concentration and the selected orbital configuration was unchanged from modern.

While this certainly is within the range of reconstructed values, presumably it lies near the upper part of that range? How do the models perform in the middle of that range or at the lower end? If this study has systematically biased the forcing to the upper range of values and gotten roughly the right answer, this implies that the models are systematically biased to be too insensitive, right? Either way, it's worth a discussion. Also the global mean TOA and surface energy budgets must be presented in a Table and discussed when covering this material. Also, how were aerosols handled?

- **2976:** No direct relationship between the magnitude of Pliocene SAT anomaly and Climate Sensitivity is seen.

Perhaps, or perhaps a global mean residual in the energy balance remains and it must be accounted for.

In this section the global mean E and P distributions should be presented in a Table and shown versus global MAT and compared with standard scalings (what is the % increase as a function of MAT).

- **2977:** The changes in global precipitation in Experiment 1 are dominated by the increases over the land, whereas the specified increases in SSTs are associated with very little increase in precipitation over the ocean. In Experiment 2, precipitation rates increase further to ~0.07 to 0.18mmday.

To me, this points to TOA/RESSURF issues as described above. Best to check if this is all just because evaporation is erroneously closing the energy budget..

- **2978**:In the same regions where the land/sea mask was altered (i.e. West Antarctica, the margins of East Antarctica 10 and the Hudson Bay),the 2sigma exceeds 8°C. Such high inter-model diðerences are attributable to the application of either the PlioMIP preferred or alternate experimental design (Haywood et al., 2010, 2011).

This shows how making an apples to oranges comparison increases experimental uncertainty by inflating the 2 error, but without having any real relevance to the model-data discrepancy.

- **2980:** confirms these basic trends, whilst highlighting regions of greater or lesser consistency between the model results. In the North Pacific, the SST anomaly is large (up to 5°C) and the standard deviation is generally no greater than 2°C (Fig. 3). In contrast, the SST response in the North Atlantic is weaker (2 to 3° C), and at the same time the **2σ** from the ensemble is large (locally exceeding 4° C).

This very clear pattern as described in the text is lost once aggregation is done in Figure 6. This is a weakness of the aggregation in Figure 6, not in the text. In the North Atlantic, the models are systematically biased to be too cool with respect to the data (large bias) and they have a widespread (poor reproducibility). Normally in statistics this means a poor performance on both counts, a bias in the mean and a lack of reproducibility. Unfortunately the way that the statistics are handled in this study, have models that are all over the place (but never in agreement with the data) somehow counts as a positive point in their favor (by increasing the 2 zone of uncertainty). This is a strange way to handle error.

- **2981**:In this region, 5 Experiment 2 predicts a larger anomaly in precipitation rates (wetter) over the oceans than Experiment 1. Conversely, the Experiment 1 anomaly is greater in the tropics over land (drier) than Experiment 2 (Fig. 4).

This suggests to me undiagnosed differences in the implied freshwater ocean+atmosphere transports plus potentially differing global mean residuals. This issue is touched lightly upon here:

- For Experiment 1, and to a lesser degree Experiment 2, the MMM differences in mPWP climate are closely linked to the specified boundary conditions provided by the 15 PRISM3D data set.

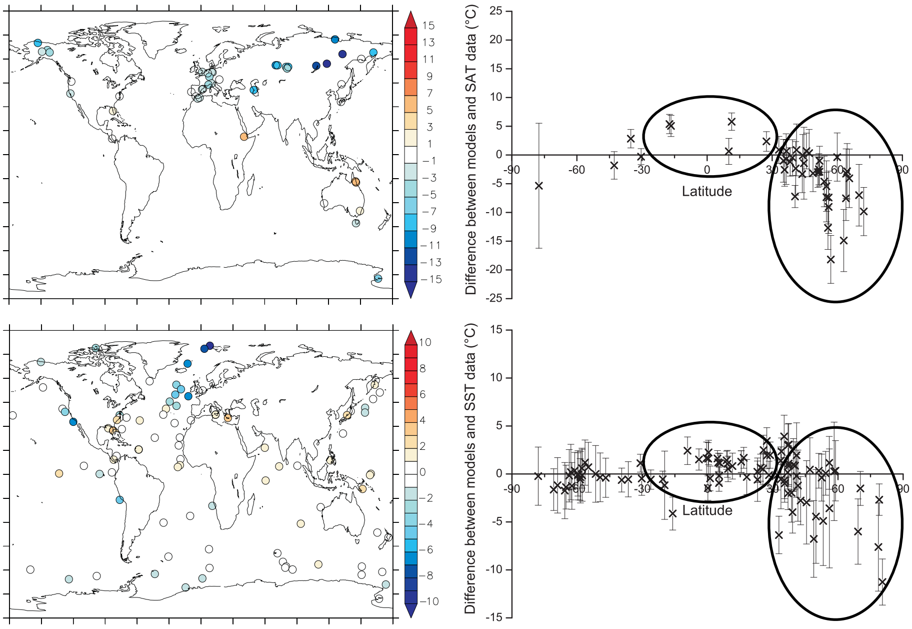
Altered SST patterns, sea and land ice volumes are a first order control on the simulated variations of the mPWP climate relative to the pre-industrial. The variations in climate are driven by changes in sensible and latent heat fluxes (SST driven), and variations in ocean/ atmosphere heat exchange caused by differences in sea ice.

But, the point needs further explanation and perhaps some diagnostics for those for whom this is not all intuitively obvious.

- **2983**: The analysis shown in Fig. 5 demonstrates a broad concordance between data and models apart from in the Northern North Atlantic and Nordic Seas. Here the MMM underestimates the magnitude of change by as much as 8 to 10 ○C. The calculated 2σ on the MMM SAT and SST anomalies indicates that the majority of the discrepancies between model results and proxy estimates are not statistically significant to a 95 % confidence interval.

The underlying statistical model here is as clear as mud. Maybe I'm confused and ignorant, but maybe the average reader is as well, so please bear with me. Is it assumed that the errors are systematic or random, and if random, what is the underlying random model? Where are the confidence intervals coming from? Are you assuming normal distributions of error, and if so why? Why are errors not adding in quadrature? How is spatial co-variance being treated? How many independent sample are there? Why aren't you aggregating and getting robust means? When I look at Figure 5 I reach a completely different conclusion than the authors.

Here's the Figure



What I see are on land in the extratropical Northern Hemisphere is approximately 27 values with negative anomalies, and 7 values with nearly 0 anomalies, and no values with positive anomalies. One can use whatever statistical test one wants, but there is no way that this is not a significant bias. Including ocean values in that analysis would not change the outcome one iota since it would add only a couple of warm bias points and a raft of cold bias points.

In the tropics the opposite pattern occurs (from 30°N to 30°S on land and sea) I count 14 strongly positively biased points, 3 strongly negative biased points, and 12 neutral points. Of course I'm just eyeballing this, but again, any reasonable statistical analysis would crunch those numbers into a large positive bias.

I suggest l binning into tropical and extratropical bins as described above and a better way of aggregating the statistics to constrain the errors in temperature gradient.

- **2986**: equilibrium state of a world at 405ppmv of CO2. To convert this to the usual definition of ESS (i.e. a CO2 doubling from 280 to 560 ppmv), the Pliocene warming is multiplied by ln(560.0/280.0)/ln(405.0/280.0) = 1.88.

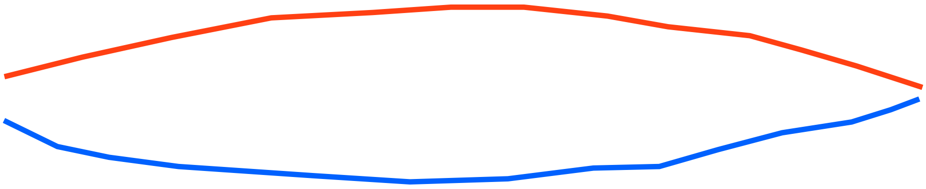
If global TOA is nonzero this effect must be calculated by adjusting the temperature change to the one it should have after TOA has reached zero. Also worth mentioning at some point Hansen's efficacy concept since it appears relevant here.

- 2987:The marine point-based DMC shown in Fig. 5 demonstrates that even in the region 20 where the proxy-derived SST anomalies are at their greatest, the 2σ calculated from the PlioMIP ensemble makes it difficult to attribute statistical significance to the vast majority of site by site data/model mismatches at a 95 % confidence level.

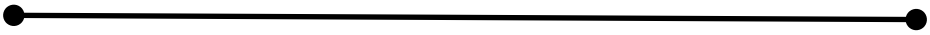
Same problems as described above (where does 95% come from? How are systematic biases in the mean with wide error being considered?). The authors are incorrectly equating enhanced intermodel spread as feature of the models rather than as a deficiency.

Consider the following analogy. I assign 50 students to measure a known standard 50mg weight on scales. The students produce values that range for 50mg to 1kg. The students should get points off both because of a mean bias and for a lack of (ensemble) reproducibility. The fact that the spread is huge and they get the wrong mean is a failure of two kinds. In this example the 2σ variation might easily be 250 mg--using the methodology in this study the students would be doing great! Currently the authors are incorrectly using the wide spread of the models as a way of reducing the model-data mismatch when in fact the models may all biased and have a large noise associated with them.

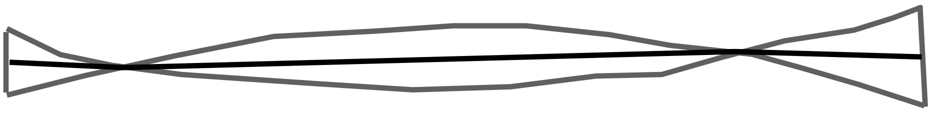
By performing this analysis using the MMM and using the standard error in this way one can achieve model-data match to 2σ in a trivially and misleading way. Let's imagine the following two model simulations:



and let's assume for the sake of simplicity that the proxy data distribution was:



Then the MMM plus variability based error bar might look something like this.



Needless to say that MMM plus the intermodel-variability based error envelop matches excellently with the proxy data when analyzed in this way. Of course, this is a specious result that arises by taking two *distributions that are completely different than the proxy data, averaging them and adding what should truly have been errors in such a way that they actually make it easier for the model and data to match.*

Clearly this is a strawman argument, but the basic point remains. The MMM analysis performed in this study can be misleading and I think the weaknesses of the technique are having important influence on the conclusions reached. For example, the 2σ-values on the upper range of temperatures in the Northern high latitudes that come close to matching the proxy records may come from models that are simply too warm everywhere (i.e. from the same simulations that are much too warm in the tropics). A proper skill metric gives negative scores to inaccurate predictions and random (guessing) predictions. The metric used in this study incorrectly puts systematic model biases into the column of model positive skill (by reducing model-data mismatch through artificial inflation of the standard deviation).

- **2987:** difficult for proxy-data or climate modelling to meaningfully inform the other regarding performance, until uncertainties in the reconstruction as well as modelling of Pliocene warmth are better quantified and then reduced. However, much of the signal of data/model discord in the Northern Hemisphere is not significant at a 95% confidence interval, and this conclusion is drawn before uncertainties in geological proxies are included.

Sorry to be repetitive, but I think that the paper currently ignores a large and robust signal of model data mismatch. The methods here might be suitable if the main point of the paper was that mismatch was large (because the statistical methodology is biased to be overly lenient to the models), but currently the main conclusions of the study potentially arise through the use of a biased estimator and are therefore not well supported.

Figure 6.

Please indicate what regions are averaged over in Figure 6, the regions do not appear to capture the major centres of action of the data. I suggest, in addition to what is done here that alternative binning might produce a more accurate diagnosis of robust model-data differences.