Response to the comments of S.J. Phipps

We would like to thank the reviewer for his detailed comments on our manuscript and his positive evaluation. Our responses to his comments are presented below in blue italic fonts while the comments themselves are in normal font.

Overview

This manuscript is a product of the joint PAGES2k-PMIP3 workshop held in Madrid in November 2013. It makes use of the PAGES2k Consortium continental-scale temperature reconstructions and a suite of transient climate model simulations of the last millennium, particularly those conducted according to the CMIP5/PMIP3 protocol. This dataset is used to fulfil three aims: to assess the consistency between the model simulations and the reconstructions; to investigate the links between different regions; and to assess the role of external forcings in driving reconstructed temperature changes. A range of state-of-the-art statistical methods is employed.
In the Northern Hemisphere, there is reasonable agreement between the models and the reconstructions. External forcings are found to be important in driving temperature changes. In the Southern Hemisphere, however, the models appear to over-estimate the response to external forcings relative to the reconstructions. As a result, the model simulations exhibit a greater degree of spatial coherence than the reconstructions.
Overall, the manuscript represents an impressive body of work. The analysis is thorough, and the presentation of the results is clear. By comprehensively comparing a multi-model ensemble with a suite of reconstructions at the continental scale, the manuscript represents a significant advance in our understanding of the climate of the last millennium. I concur with the statement by the authors that "our results for the Northern Hemisphere suggest a convergence of our understanding of climate variability over the past 1000 years, but there remain many open questions for the Southern Hemisphere". The manuscript is entirely suitable for publication in Climate of the Past, and I recommend that it be published subject to the following comments being taken into consideration.
In the interests of transparency, I wish to state that I was one of the participants in the PAGES2k-PMIP3 workshop. I saw some early versions of this manuscript, as these were distributed to the workshop participants. However, I did not make any contribution towards the writing or development of the manuscript, and therefore I do not see any reason why I should not act as a referee.

General comments

1. Length

The manuscript is relatively long: 73 pages (in the discussion format) plus a 28-page supplement. Generally, this length is justified by the scope of the manuscript. However, I feel that the clarity of the text could be improved by tightening it in places, particularly in Sections 1 and 3. I would therefore encourage the authors to edit the text for length.
In specific regard to Section 3, which is currently nine pages long, I feel that much of this information could be transferred to the Supplement. Rather than providing extensive descriptions of each technique in the main text, I suggest that Section 3 could be restricted to 1-2 paragraph summaries that briefly summarise each technique and the motivation for using it. The interested reader could then consult the Supplement for more detailed information.
The manuscript also contains some excessively long paragraphs. This can impede readability, particularly as the longer paragraphs can lack a sharp focus. As an example, the paragraph that runs from line 9 on page 2490 to line 4 on page 2491 discusses the Southern Hemisphere, modes of climate variability, inter-hemispheric linkages and finally the Northern Hemisphere. I am afraid that the point that this paragraph is trying to make is unclear to me, and I feel that much of it could simply be removed. The manuscript would benefit from abbreviating longer paragraphs such as these, or from breaking them up into multiple shorter and more focussed paragraphs.

In order to reduce the length of the manuscript, we will suppress Section 3. A part of the material will be included in Section 4, as a short description of the techniques, before the diagnostics are actually presented. The remaining information, if required, will be placed in the supplementary material. Furthermore, the long paragraphs will be split in shorter, more focused ones and/or reduced in length, in particular the one that runs from line 9 on page 2490 to line 4 on page 2491. The manuscript will also be reduced wherever it appears possible while still keeping the text comprehensive, precise and clear.

2. PMIP3 versus non-PMIP3 simulations

The title and abstract of the manuscript states that it studies the PMIP3 simulations, and they are referenced as such throughout the text. The value of protocols such as that developed by PMIP3 is that they enable the inter-comparison of model simulations driven by consistent external forcings. However, the selection criteria outlined in Section 2.2 differ somewhat from the PMIP3 protocol. In particular, orbital forcing is not applied as a selection criterion despite the fact that it is required by the PMIP3 protocol and can be an important forcing on the continental scales studied by the manuscript. Non-PMIP3 solar reconstructions are also permitted (particularly CESM1, which uses a solar forcing timeseries in which the amplitude of variations has been artificially doubled). While there is considerable uncertainty in reconstructions of solar forcing, this nonetheless strikes me as undesirable. Furthermore, some of the PMIP3 simulations that do meet the stated criteria are excluded from the manuscript.

The authors should provide further justification for their specific selection of models, including: (i) why they did not simply select models on the basis that they follow the PMIP3 protocol; (ii) why some of the PMIP3 simulations were excluded; and (iii) why the CESM1 and COSMOS simulations were included, but not other non-PMIP3 simulations such as those studied by Fernandez-Donado et al. (2013). In regard to (ii), I believe that there were valid technical reasons for this, such as excessive drift, incorrect application of forcings, or deviations from the PMIP3 protocol; this should be explicitly stated. If the authors wish to retain the non-PMIP3 simulations, they should also make appropriate changes to the text (for example, removing "PMIP3" from the title and refraining from referring to the simulations as "PMIP3 simulations").

We appreciate the referee's concerns that the use of the term "PMIP3 simulations" should ideally refer to all and only all the simulations that comply with PMIP3 protocol. We have applied a more loose definition here. A first goal was to remove some simulations that have known biases. A second one was to include additional simulations whose design was close enough to the PMIP3 protocol to provide a meaningful comparison. We will specify this more explicitly in the revised version of the manuscript. Despite this limitation, we suggest to keep reference to "PMIP3" for simplicity and for an easy reference in some parts of the text (in particular in the title) while removing it when possible and explain very clearly in the manuscript, starting in the abstract, that we analyze 9 simulations covering the pre-industrial millennium and the historical period, 7 of them following (or being compatible with) the PMIP3/CMIP5 protocol and two having an experimental design that is close to one required by PMIP3.

Furthermore, in the revised version of the text we will more precisely describe the selection criteria and why we included pre-PMIP3 simulations, answering the referees' questions:

(i) why they did not simply select models on the basis that they follow the PMIP3 protocol; (ii) why some of the PMIP3 simulations were excluded:

For the reconstructions-simulation results comparison we defined as a prerequisite that the records include the pre-industrial millennium (850-1849) as well as the historical period (1850-2000). This is important to judge how the simulations reproduce the transition from the weak forcing in the "past1000" simulations to the relatively strong aerosol and greenhouse gas forcing in the "historical"

runs. Unfortunately, the PMIP3 protocol did not include this requirement and the "official" "past1000" period had been defined as 850-1849. Therefore some modelling groups followed the protocol and did not provide matching "historical" simulations, or in some cases, the ESGF data-base and further inquiries with the modeling groups did not allow us identifying these (e.g. bcc-csm1-1). Moreover, initial analyses and cross-checking with the generators of the simulations revealed unacceptable drift in some simulations These experiments were excluded, (in particular the one performed with MIROC) since a simple detrending (as applied for the GISS model simulations which show a moderate drift) would have unknown consequences at regional scales and thus potentially a large impact on our analyses.

(iii) why the CESM1 and COSMOS simulations were included, but not other non-PMIP3 simulations such as those studied by Fernandez-Donado et al. (2013):

Many of the pre-PMIP3 simulations analysed in Fernandez-Donado et al. (2013) ended before 2000 and thus did not cover the period required for our investigations. Furthermore, the amplitude of solar forcing during the millennium (see IPCC-AR5, Chapter 5; Fernandez-Donado, 2013) in those simulations is often much larger than in the PMIP3 simulations. Nevertheless, we decided to include simulations with different magnitudes of solar forcing since many uncertainties remains on this issue, as also discussed in the framework of the PMIP3 protocol (Schmidt et al., 2012). In particular, it is one important finding of the paper that simulations with stronger amplitude in the TSI do not give a systematically better agreement with the continental-scale reconstructions, confirming previous work on the subject at larger scale. However, we restricted, the "high-solar" simulations to the most recent pre-PMIP3 simulations (COSMOS complies with the PMIP3 protocol in all other aspects) and the CESM1 simulations by Lehner et al. (2015) as they apply an amplitude of the forcing that is still much lower than in many simulations analysed in Fernandez-Donado et al. (2013), they use model versions that are closer to the ones used in experiments following PMIP3 protocol and thus more directly comparable and, as mentioned above, they cover the full period investigated here. Finally, CESM1 simulations indeed used fixed orbital parameters but Lehner et al. (2015) show that this do not affect the simulated regional climate (e.g., pan-Arctic land temperatures) in a detectable fashion, meaning that it has at best a very moderate impact on the comparisons performed in the manuscript.

Specific comments

P2486, L20-23: This sentence could also acknowledge that, because of computational constraints, models have limited spatial resolution and are deliberately incomplete (i.e. they can omit known processes, such as the carbon cycle or atmospheric chemistry).

This will be added in the revised version

P2488, L2: I think "cannot be explained by" should simply be "is due to" or similar; also, remove "changes in", as "forcing" already refers to a change in boundary conditions.

As suggested, "cannot be explained by" will be replaced by "is not due to". Besides, to our experience, there is no general consensus in the usage of the term 'forcing', which is sometimes used instead of boundary conditions or of changes in boundary conditions. We thus prefer to use 'changes in forcing' to be explicit and avoid potential confusion.

P2488, L3-5: This sentence blends two distinct issues, and could distinguish more clearly between them: the fidelity of the representation of internal variability within a model, and the specific instance of internal variability within an individual simulation.

We agree that this sentence contained too much information to be clear. As our goal in the introduction was not to discuss in detail the origin of the different realizations of internal variability between the models, we propose to simply remove the sentence of the revised version as explaining them in details will make the paper even longer.

P2489, L1-2: Also PAGES 2k Consortium (2013).

The reference will be added in the revised version.

P2493, L15: Please clarify what is meant by "They": procedures, uncertainties?

It is the uncertainties. This will be specified explicitly in the revised version.

P2494, L13: The variable name "tas" is enforced by the Climate Model Output Rewriter, but the name was chosen by CMIP5.

We will add 'of CMIP5' in the revised version of the manuscript to be precise.

P2494, L24-26: I'm not sure if this is strictly correct: when studying power spectra, for example, stochastic internal variability would not prevent perfect data-model agreement.

We agree on this point. We will thus add in the revised version that this is valid when directly comparing time series.

P2495, L1-6: It would be helpful if the authors could more clearly link the specific combination of methods chosen to the three aims of the manuscript.

This will be done in the revised version of the manuscript.

P2496, L18-22: This sentence is unclear to me; please clarify. Section 3.1: Further to my comment on Section 3 above, this section is a good example of where a concise 1-2 paragraph summary would be better suited to most readers than an extended technical description. Most of this text could happily be transferred to the Supplement.

As suggested by the reviewer, most of this section will be transferred to the supplement.

P2499, L14: Is this the correct reference? And, if so, why was this reconstruction used and not one or more of the reconstructions that were used to drive the models?

We are sorry that the paragraph was misleading. The composites of the solar forcing are focused on time intervals of low solar activity following those selected in the IPCC AR5 (Fig. 5.8, Masson-Delmotte et al., 2013) for the sake of a better comparison. The text will be changed to make this point clear.

P2504, L5-6: This is not a genuine like-for-like comparison, as some of the reconstructions are land/ocean and some are land-only.

The reviewer is right. This will be specified in the revised version of the text.

P2510, L16-17: "very high" is a value judgement; very high relative to what? It would be better to use a more neutral statement such as "higher than +0.5".

This will be done in the revised version.

P2510, L27-P2511, L1: What is the basis for this statement? Visually, the correlations for the full period appear to be only slightly higher than for the pre-industrial period. This suggests that forced variations during the pre-industrial period are the dominant explanation of the strong correlations.

The correlation for the industrial period are generally higher than for the pre-industrial period (see also Figure 7) making the correlations higher for the whole period than for the pre-industrial period. However, we agree that the correlations are also generally high for the pre-industrial period. We will thus modify the text in the revised version to make this clearer, insisting that it is the simulated high correlations for the last century that are likely to be a consequence of the rather homogeneous and strong anthropogenic warming in the simulations.

Section 6.1: (i) I have doubts regarding the methodology used to convert the volcanic reconstructions into timeseries of radiative forcing. The authors apply an established scaling factor of -20Wm^{-2} to convert aerosol optical depth into radiative forcing. However, I believe that this scaling factor is only valid for global means. Regionally, the appropriate scaling factor will depend upon the shortwave radiative fluxes; a given aerosol optical depth will scatter a given fraction of shortwave radiation, but this will equate to a greater radiative forcing in the tropics than at the poles. I can see no obvious solution to this, and I suggest that the radiative forcing data simply be removed from this section. This would not prevent the data-model comparison, which does not depend upon the radiative forcings anyway. (ii) When comparing the regions, the authors should comment on the fact that some of the reconstructions are land/ocean and some are land-only. For example, Europe and Asia, which have the strongest volcanic responses, are land-only reconstructions.

We agree with the reviewer that expressing the volcanic forcing at regional scales is quite uncertain. In the submitted version, our goal was to give an order of magnitude of the forcing but as it can be

misleading we will remove it in the revised version as suggested. For the point (ii), a comment will be added to explain that the different magnitudes of the responses may be due to the fact that some regions includes relative large ocean regions while others are land-only.

P2523, L20-27: Agreed! From this perspective, it would be helpful if the manuscript could document the implementation of volcanic forcing within each model e.g. in Table 1 or Supplementary Table S1.

Documenting precisely the implementation of the volcanic forcing would indeed be helpful but we consider that it is out of the scope of our study. In order to be accurate, this can be relatively long for some models that include a sophisticated representation of the radiative effects of volcanic aerosols while the link between the detailed implementation and the response of the model (which is the subject of our analysis) will likely be not straightforward.

Minor edits

Unless specified below all the minor edits will be included in the revised version.

P2484, L9: Insert "the" before "time".
P2484, L16: latitudes -> latitude.
P2485, L3: Remove "the".
P2485, L4: Constraints on, and uncertainties in, external :
P2485, L9: places -> place.
P2486, L5: "observing" -> "evaluating" or similar?

The suggestion appears not clear to us but the sentence will be modified hopefully in a satisfactory way.

P2486, L14: "univocal" -> "unequivocal".
P2486, L24: Remove "also" before "not"; insert "either" after "understood".
P2490, L6: Insert "that" before "there".
P2490, L8: "2014" -> "2013".
P2491, L8: Insert "the" before "PAGES".
P2495, L2: "as" -> "to".
P2495, L13: Insert "by the fact that" after "and".
P2495, L25: "exchangable" -> "statistically interchangeable".

"Exchangeable" is the standard term (see for instance Bothe et al. 2013) so we prefer to keep it.

P2496, L12-13: Insert "the" before "simulation" and "target".
P2496, L14: "too" -> "excessively".
P2497, L2: "exchangeability" -> "interchangeability"?

*"*Exchangeability*" is the standard term (see for instance Bothe et al. 2013) so we prefer to keep it*

P2499, L1: "Untermann" -> "Unterman".
P2500, L3: Insert "a" before "tool".
P2501, L10: "in" -> "over".
P2502, L16: Remove one of the instances of "relative".
P2504, L4: "estimating" -> "calculating"?
P2504, L5: "control runs" -> "pre-industrial control simulations".
P2504, L12: Insert "the" before "CESM".
P2504, L15: "but" -> "and".
P2504, L19: Insert "than the models" after "variability".
P2510, L20: Remove "already".
P2510, L22: "do" -> "does".
P2514, L19: "Unternman" -> "Unterman".
P2515, L22: "done" -> "performed".
P2516, L29: Remove "ensembles"?
P2517, L19: "hamming" -> "Hamming".
P2519, L8: "constrains" -> "constraints".
P2521, L4: Insert "the" before "regional".
P2521, L7: Insert "the" before "continental".

P2523, L3: “yield” -> “yields a”.
P2523, L5: Insert “the” before “hemispheric”.
P2525, L23: Insert “the” before “continental”.
P2526, L2: Insert “a” before “significant”.
P2526, L17: Insert “the” before “basis”.
P2546, figure caption: Insert “the” before “full length”.
Supp P1, L40: “Untermann” -> “Unterman”.

Response to the comments of Referee #2

We would like to thank the reviewer for his/her positive evaluation of our work and for the constructive comments. Our responses to his/her comments are presented below in blue italic fonts while the comments themselves are in normal font.

This manuscript presents an analysis of the performance of climate model simulations of the last millennium. I don't really have much to say in the way of suggestions and recommend it for publication in CP. The analysis is extensive and seems appropriate, the paper is not earth-shattering in conclusions, but it is certainly a useful analysis of the current state of climate model performance (as far as can be deduced from recent proxy-based reconstructions). It is a lengthy paper but as the model-data comparison is extensive, this is not entirely avoidable. However, the English in some places is a little hard work, and it could do with some editing and rewriting in parts.

As also suggested by the first referee, we will edit and rewrite some parts of the text for the revised version to make it clearer and easier to read. A part of the material will also be suppressed or moved to the supplement to reduce the length.

One area that particularly stands out is the description of volcanic response in 6.1, where it is not always clear which two quantities are being compared for many instances of "larger" and "smaller" etc. In addition to the overall magnitude of forcing, I would expect that the regional distribution of the forcing could be somewhat uncertain here (and thus lead to regional discrepancies between models and reconstructions) but this does not seem to be mentioned in the text.

In the revised version, we will ensure that when two quantities are compared, it is clear which one is larger or smaller. As the magnitude of the forcing is indeed very uncertain at regional scale, we will remove it from the plots (Fig. 8; Figs S11-S13) and discuss the uncertainties in this forcing in the revised version of the manuscript.

I'd also prefer more direct language that does not skate around the issue of model data disagreements where they exist. Throughout the paper, discrepancies are often attributed to "uncertainties" when in fact uncertainty, if correctly accounted for (which the methods used can potentially do), should not in itself give rise to significant discrepancies. The problem is surely with errors that lie outside the range of estimated (or tested) uncertainties. The authors could do worse than globally search for "uncertain" and ask themselves whether it would not be clearer to talk frankly about errors.

The paper also seems to take a rather rosy view of the reconstructions. I realise that the authors did not set out to assess the reconstructions, but taking them as a ground truth (albeit with their stated uncertainties) seems potentially misleading. Given the wide range of results reported for NH temperatures e.g. in the IPCC AR4 "spaghetti plot" of Fig 6.10 (which shows persistent disagreements of order 0.5C even after heavy smoothing), it seems optimistic to expect reconstruction accuracy to be reliable on regional scales. Where models agree reasonably with each other (but not with the reconstruction) and forcing uncertainties are not considered to be large, gross errors in reconstructions cannot be excluded.

We agree with the reviewer that it is important to make the distinction between uncertainties and discrepancies. We also agree that we cannot exclude gross errors in reconstruction that are larger than the given uncertainties. Nevertheless, it is impossible to state for sure that a difference between reconstructions and model results is due to errors in model physics, in the forcing or in the reconstructions. This is why in that case we prefer to mention the disagreement and then discuss its

possible causes. The fact that the quality of reconstructions is a potential source of discrepancy between model results and reconstructions (in particular in the Southern Hemisphere because of the lower number of records compared to the Northern Hemisphere) is mentioned several times in the submitted manuscript, starting in the abstract. Nevertheless, we will check again carefully in the revised version, as suggested, that in all sections we are clear enough about those points, stating it explicitly each time when a difference between reconstructions and model results cannot be attributed to uncertainties alone and discussing all the potential sources of disagreement, including inadequate or underestimated uncertainties in the reconstructions. To better elucidate the scope of the uncertainty estimation for reconstructions, we will be more explicit about which sources of uncertainty were accounted for in each regional reconstruction.