Clim. Past Discuss., 9, C1235–C1258, 2013
www.clim-past-discuss.net/9/C1235/2013/

Climate of the Past Discussions

Interactive comment on "Using paleo-climate comparisons to constrain future projections in CMIP5" *by* G. A. Schmidt et al.

G. A. Schmidt et al.

gavin.a.schmidt@nasa.gov

Received and published: 26 June 2013

We thank both reviewers for their constructive comments on the manuscript. Judging from both responses, it is obvious that our motivations for this paper were not clear, or rather were perhaps too ambitious for the material at hand. We will clarify that in the revision.

In the examples we used, there were some cases where we had a robust relationship in the models for past and future, cases where some models match the paleo-data better than others and give distinct predictions, cases where the constraints fail because of lack of data, or dependence on uncertain forcings, or an incoherence between past and future diagnostics. Our point in using all of these was to be clear that deriving useful constraints is not trivial and that there are pitfalls that can be detected and thus

avoided in future work. We respond more specifically to points raised, below.

Anonymous Referee 1

The manuscript seems to be a follow-up report of a workshop held in Hawaii in 2012 on joint analysis of paleo simulations and proxy reconstructions.

While this paper does derive from this workshop (which was explicitly stated in the acknowledgements), the paper is not a workshop report, rather an attempt at a synthesis of discussions and examples initially presented there and then expanded.

The abstract and the introduction prompt the reader to expect a theoretical framework and best-practice recommendations on how to use proxy records to analyse the skill of paleo model simulations, and how the joint comparison of simulations and reconstructions can lead to better constrain climate projections. (this last point is prominently included in the title).

This is the aim, though we realise we may have fallen short of the task. A clearer statement of aims would be to produce an exploration of the issues that arise in attempting to constrain the future projections from paleo-model data comparisons. This is inevitably an incomplete picture since the raw data that allows this to work (the CMIP5/PMIP3 archive) is still a work in progress and has not been fully explored.

The manuscript is generally clearly written and I think it can be understood by both the modelling and proxy community. However, I was disappointed by an unclear manuscript structure and the quite disjoint sections.

This point is echoed in both reviewers comments and will be improved in the rewrite.

After reading the manuscript, I was left wondering about which the theoretical framework and the best practice recommendations actually were. I am aware that to write a workshop report for a scientific journal is not always easy, as the objectives of the workshop participants are quite often too diverse. But then the manuscript should be clear at the outset that this text is not intended to be a proper research article but

rather a workshop report with some working examples without no specific results on this matter.

While we agree in principle, we are not aiming to write a workshop report. Rather, we do intend this to be a synthesis of research directions and pitfalls using examples from the workshop to illustrate them.

After a general introduction, the manuscript includes a collection of section that have very little connection to one another, and that perhaps due to space limitations, are only a short introduction to a line of research without going into deeper detail. Some of these sections included in the manuscript are clearly more useful than others.

This is very much a subjective statement. However, we will be clearer in the rewrite what the use of each section actually is.

For instance, section 4.2 on how to constrain climate sensitivity from simulations of the Last Glacial Maximum, or section 3.1 showing that the temperature signal of the external forcing may be more strongly felt in the tropics than at high latitudes or over the continents, and thus past temperatures in these regions may be a better predictor for future temperatures (Here, however, I would argue that the magnitude of the response is only one aspect of the metric, the other aspect being the magnitude of the internal variability).

We are not quite sure what scale of internal variability is being alluded to here. The LGM proxy-based datasets are multi-millennial averages, and the models are in quasi-equilibrium. This is therefore a coherent comparison. However, we do not have good global or regional proxies for shorter term variability for this time period, so we are not sure how those aspects could be usefully compared to model output. We acknowledge that changes in internal variability could actually have an impact on climate proxies and this is a topic that will be investigated in the future with the help of forward modelling of proxies. This idea could be added in the new version of the manuscript if this is what the reviewer meant. Internal variability on longer timescales (i.e. D/O Heinrich events)

almost certainly involve ocean/dynamic ice-sheet interactions and so are beyond the scope of the CMIP5 models (which impose ice sheets as a boundary condition).

Another enlightening example is the simulation of ENSO (section 5.1) although in this case the authors make the point that the modelling results are too ambiguous as to provide a useful metric for future projections.

We agree that this is an important point. However, it may be that improvements in proxy data and models may change this in future.

I did not find other sections particularly useful. For instance, I fail to see the point of section 5.2 on the spectral properties of the simulated temperatures and proxy records. How can the spectra be used to quantitatively discriminate among models ?

If the spectra of different models were different and we had spectra from the paleo-reconstructions, we don't see why *in principle* this comparison could not be determinative of which models were better. The point of this illustration is that the uncertain forcings (which are an input to the models) dominate the variation in the spectra, precluding an easy use of this metric to distinguish between models. We think this is an important point to make, especially since many previous paleo-model/data comparisons have not fully explored the uncertainties in forcings. Other reasons for a mismatch - biases in the proxy reconstructions or missing feedbacks and variability in the models - can't really be addressed if the dominant reason for variance are the input forcing time-series.

This section seems to be rather a short summary of the deficiencies of paleo simulations in replicating different spectral regimes, and the authors themselves do not seem particularly convinced judging from the caveats about forcing uncertainties added at the end of this section.

Our text tried to make clear that *since* the different spectral regimes are obviously affected by reasonable uncertainties in the forcings, that implies that one is currently

unable to constrain the models by comparison to the reconstructed spectra. The example is one where the ensemble of models with different choices of the forcings was essential in preventing an erroneous conclusion being drawn.

Finally, section 5.3 appears as short discussion about the risks of using the classical Palmer Drought Severity Index as an indicator of past or future droughts, but it does not contain any indication of how the frequency or intensity of simulated droughts may serve as a discriminating metric across model simulations.

Agreed. It nowhere claimed to. Rather, it is an exploration of why some metrics may fail to provide future constraints, despite the existence of paleo-reconstructions.

This section also misses an important discussion on what is the variable that can be effectively reconstructed from proxies. Is it soil moisture, is it precipitation, is it a mixture of both and if yes, how could they be disentangled ? is it some modified Palmer Index? If this section were to be coherent with, say, section 4.2, the reader would expect suggestions of where and in which seasons should past droughts be used to constrain future drought projections. As it is, the section is narrowly focused.

Yes. But it is an illustration of what can happen, not a comprehensive search for something that might work. We are hopeful that people will explore that in future.

It is not straightforward to see how the recommendations included in the Conclusions can be derived from examples of the previous sections, although they all sound reasonable. To me they are probably the result of general discussions held at the workshops, but they certainly are not connected to the specific examples shown previously.

These are general points for which the examples shown are illustrations. We will make the links more obvious in the rewrite though.

Thus my general recommendation would be: -to warn the reader at the outset that the manuscript is a workshop 'progress' report that includes some ad-hoc suggestions, but certainly it does not include a theoretical framework or best-practice suggestions.

We disagree - 'best practice' suggestions are often a little ad hoc and there is discussion of a number of (relatively informal) theoretical frameworks. However, it is perhaps clear that our conception of what counts as a theoretical framework is arguable, and so we will adjust this language in the revision. We agree that this is a work in progress.

As a workshop report, I think the manuscript may be useful and informative. I would not say it fulfils the standard of a classical research article or review paper.

We do not aim to write a workshop report, but we hope that the revised manuscript will fulfil the reviewer's standards more fully.

-to homogenize the different sections, not so much in terms of style, but in terms of content and scope. They are now widely different.

They are. But they are illustrating different points. We will rewrite the text to improve homogeneity.

The present version looks rather like a juxtaposition of different texts written individual authors, and some of them are not necessarily related to constraining future projections from paleo simulations. This is in my view particularly true for sections 5.2 and 5.3.

Fair point and we will address that. The points in 5.2 and 5.3 are more subtle, and point to cases where no clear constraint is possible. We think these counter examples are useful in demonstrating limitations that other researchers will undoubtedly encounter. We will clarify these points in the text.

I think that this over arching goal should be clearly discernible in all examples, even though it can be difficult to provide a detailed derivation for each of them.

We will definitely be clearer in expressing them.

Some particular suggestions: 'The scale over which a record is representative can be a major issue in comparing paleodata and model output. All types of records are responding to local conditions, and for basic meteorological variables it is rare for a

record to be representative for spatial scales of more than 50–100 km (though many records, such as tropical ice core d18O, may have strong correlations to climate further afield; e.g. Schmidt et al., 2007). Comparisons at these scales often require some form of dynamical or statistical down-..' The spatial coherence depends in general on the time scale. Usually the spatial coherence strengthens with time scale, at least for temperature. It is unclear to me to which time scale the quoted 50-100 km would apply.

This is a good point and will be addressed in the rewrite.

The manuscript stress the advantage to use the same climate model for paleo simulations and future projections so that the lessons obtained from past climate can be better projected into the future. However, in the CMIP5 data base, this is not always the case. The resolution/version of the paleo models is often different from those used for climate projections. To what extent they can be considered the same model is an open question.

Actually in CMIP5, it is the norm that the models are in fact using the same physics at the same resolution. Indeed, it is exactly this feature that makes this exercise worthwhile. We agree that two distinct resolutions using the same basic model code should not be considered the same model, and we do not use any such cases in our examples. We will add this statement explicitly to section 2.4.

Referee 2 (Tamsin Edwards)

First, my sincere apologies for the delay in posting this review. And apologies if, in my aim to post this slightly more quickly, I have come across as at all brusque: it was not my intention.

We all appreciate the time and effort constructive reviewing takes.

Given the author list, I hoped this would be a robust and exhaustive list of palaeoclimate relationships with future and comparisons with data.

We think that such a list might be useful, but we did not intend to create one, nor do

we think this would be possible given the limited analysis that has been done on the CMIP5 archive.

Perhaps not statistical inference – i.e. actual constraints on the future – but I did expect a useful reference and baseline with which others can start.

We did (do) indeed intend to produce a useful reference and a baseline for others. That the reviewer does not think this was successful is a testament to the work we need to do on the revision.

However the contents do not fulfil the promises of "theoretical framework" nor "best practice". It seems to be a collection of completely separate studies of palaeodata and/or models, which are often imprecisely described, with claims weakly or not supported by data, and that are not coherent with each other or the remit of the paper.

We think this is a little harsh. The examples are indeed disparate, but they are all illustrative of selected and different aspects of the problem. None of them are claimed to be definitive and so most claims are indeed exploratory. The coherence comes from the (admittedly poorly explained) scope and remit, rather than each example being applied to the exact same problem.

It is not a well-written or well-structured manuscript; it took me quite a long time to discern the scientific questions asked and the degree to which they were answered for each section. In my view the manuscript needs (a) substantial rewriting as a set of simple studies of PMIP3 ensemble characteristics, or else removal of irrelevant content and rewriting to more clearly address the stated remit (which may leave too little material); (b) removal of weakly or unsupported claims, and (c) rewriting of imprecise and unclear descriptions of work.

We will certainly rewrite anything deemed to be imprecise or unclear, and ensure that claims made are supported. However, this paper is not a set of simple studies of the PMIP3 ensemble. We do intend to rewrite more clearly to address the (slightly

modified) remit. We do not agree that much of the content is irrelevant.

For the stated aims of the paper, the questions asked for each quantity should be, in this order: 1. Is there a relationship between past and future across models? If there is: 2. What do the proxy data indicate for the past? 3. What does this imply for the future?

This is certainly the crux of the matter and most of our examples follow this already. But it does not cover the entirety of what we are discussing. Missing are considerations of whether model outputs and proxies are commensurate, how one can explore variations in the answers to each of those questions across models.

Currently, where the answer to 1 is shown to be "yes", there is often no stating of the palaeodata assessment or propagation of this to the future.

This is not a fair assessment. In most sections there is a clear propagation to the future and where there isn't, this is explained (and this will be more explicit in the revision). The propagation might not occur because of deficiencies in the models, in the proxy interpretations, in the uncertain boundary conditions etc. It is worth exploring each of these cases in our examples because they will occur in other work being done.

Where the answer is "no", this is often stated at the end of the section after (for the stated remit of this paper) irrelevant model-data comparisons are made.

Since our aim was to demonstrate issues that arise in attempting to constrain the future, it is incumbent on us to illustrate cases where it doesn't work as much as where it does. Indeed, since the cases picked don't work for distinct reasons, it is definitely worth exploring since other researchers will come across similar issues in their work. These cases then are far from 'irrelevant'. We are sure that the reviewer does not mean to imply that only positive results are publishable?

In cases where there wasn't a relationship between past and future in the models, the model-data comparisons are still key to investigating potential biases in both the mod-

els and proxy data, which may be used for further model development and exploration and improvement of these metrics.

And often (1) is not answered at all. The paper's aims are therefore not addressed, so most of the material should be adapted for a paper with the aims of comparing and understanding models rather than constraining future projections.

We are not interested in merely comparing models. Very little is learned in such exercises other than models differ. We are rather interested in exploring how and why paleo-constraints work (or not) and why.

For example, I do not understand the point of Sections 3 or 5. Neither examine relationships between the past and future, except (partially) Section 5.3.

We do not follow the reviewer's reasoning here. In Fig 1/sect. 3.1 and 2/sect. 3.2, we examine potential relationships between the lgm runs and 1and that example is included and described more provisionally. Section 5.1 and Figure 10 (lower panel) address the relationship between the 20th-century d18Ocoral trend and d18Ocoral trend projected for the future in the CMIP5 ensemble. This work provides an example of a case where there was not a relationship between the past and future response in the models; however, this model-data comparison highlights key uncertainties in the observed and simulated salinity trends within the basin and thus provides a basis for further development of the models and this potential metric.

In Section 3, for example: 3.1 Regional climate change. Relationship between past and future? No. (Not assessed directly, but e.g. GCMs rejected by Bartlein et al. span nearly the entire range for future).

We apologize for the mistake in the legend of 3.1: the Bartlein et al reconstruction is not used in this figure, on which the reconstructions are from the MARGO (2009) data set for tropical oceans and Valérie Masson-Delmotte for the compilation of temperature anomalies over East Antarctica. Nonetheless, we agree that indeed, for East Antarc-

tica for instance, the models which yield results in agreement with the reconstructions do span nearly all the possible range of results for the abrupt 4xCO2 or 1pctCO2 scenarios. But we do think this result is worth being presenting and it raises questions about what causes the range of model results to be different for lgm vs future conditions. While we cannot answer this question in the current manuscript, the question itself will be better discussed in the revised version.

3.2. Land– ocean contrast. Relationship between past and future? No. (Not assessed directly; but all models appear to have a constant ratio in both past and the future).

How is this not a direct assessment? We are able to confirm that the relationship is robust across models, data, and time period. This is actually quite impressive. This allows for a confident prediction that the relationship will exist in future climates as well.

3.3 Regional extremes. Relationship between past and future? Not assessed. (only extreme temperature versus extreme precipitation during the historical period, with a brief mention that documentary studies exist for the former).

This is a fair point, but one we address. We used this example to demonstrate future potential.

I find several of the sections do not stand up to the remit when expressed in this form. The content in Section 5 could be removed and replaced by a couple of short paragraphs for the stated purposes of the paper.

We disagree. It is useful to illustrate specific problems that can arise because of uncertain forcings or non-linear diagnostics. We often find that it is the cases that don't work that are the most useful to discuss with others.

There are also several weak or unsupported claims – not even by the standards of formal statistical inference such as goodness-of-fit testing but by simple inspection of the data, e.g.: - That there is a relationship between regional and global change for individual eras (Fig. 1): e.g. 787/19: "Such a relationship also exists for the LGM",

when the LGM data consist of a cluster of uncorrelated points with a single outlier;

The point is well-taken for the tropical LGM data in fig 1 and we will address that in the revision. It is worth stating that the results from the LGM simulations do not follow the relationship derived from the simulations of future climate for the tropics (Fig 1a and 1c) but they do for the other two regions. (North Atlantic Europe and East Antarctica). Understanding why this is so is a key question.

- That the land-sea contrasts are consistent with the LGM data (Fig. 2): the data error bounds for "NorthAtlanticEurope" are consistent with ratios from 0.9 to 3, and for the tropics from -0.1 to infinity.

We think this comment arises from the fact we did not adequately explained what the "error bars" were on Fig 2. Apologies for this. These "error bars" are actually not error bars but indicate the spatial standard deviation of the lgm vs pre-industrial temperature anomalies. This was initially plotted to compare the models' spatial variability to the reconstructions spatial variability, but this was explained neither in the main text, nor in the legend. We will update the text, figure or the figure legend to avoid this misunderstanding.

Specific Comments:

We have edited the manuscript to include all of these points where they are still relevant except for the following:

21 is dust a better example of an uncertain LGM forcing than ice sheets?

Dust is also uncertain, but it is not something that was systematically varied in the PMIP3 simulations. Some upcoming simulations will use a fully interactive dust module and this will be helpful, but multiple ice sheet reconstructions are already available and are being tested. We already know that there are important uncertainties in ice sheet topography and climate impacts are large.

p787 15 if well documented, please include a reference or two...

This shows in the Bartlein et al compilation for the continents and in the MARGO 2009 ocean reconstructions and will be added to the text.

19 relationship also exists -> don't agree! uncorrelated cluster of models plus one outlier

Yes, see the response above.

23 relationship -> both the LGM and 4xCO2 relationships rely on a single outlier. It's true they agree across experiments for the Antarctic and Europe plots, but these statements are generally too strong.

Fine, see response above

13 consistent with the LGM data -> Delete: for NorthAtlanticEurope the data error bounds would be consistent with ratios from about 0.9 to 3, and for the tropics from -0.1 to infinity. So I don't think this is worth commenting on...

See the response above. However, the inferred paleo-data lies right on the extremely robust result from the models, in multiple independent tests. We do think this is worth commenting on, though we will adjust the text for greater clarity.

27 onwards. I think you ought to make it clear that you are using about the simplest metric of drought possible, precip above a fixed threshold; drought is usually defined relative to local climatology (as you have for temperature), often include other variables such as potential evaporation and soil moisture, and may include a measure of persistence. Describe the definition of hot days in the text as well as the figure caption. I think the conclusions look rather sensitive to the point with highest hot day freq, in both EOBS and IPSL. Without the bold lines would these look like scatter? What tests were done to check the robustness of the QR results?

This will be expanded on in the text. We agree that the frequency of rainy days is a simple indicator of drought, but it does have a predictive skill on summer temperature (see Fischer et al., Geophys. Res. Lett., 2007; Vautard et al., Geophys. Res. Lett.,

2007).

The frequency of hot days is the fraction of days (in the summer) for which the temperature anomaly exceeds the 90th quantile over a reference period (1948-2011).

Quantile Regression (QR), by essence, weakly depends on the value of one single point. The slopes of QR pass standard statistical regression tests (Koenker, 2007; Quesada et al., 2012).

p790 1 describe why choose QR - from inspection of shape (e.g. funnel / triangular) of data? from a priori expectation?

Quantile regression allows one to investigate the conditional dependence of one variable to another, especially for large and small quantiles (or values).

7 of the precipitation or temperature dependence for hot or cool summers - confusing! -> of the relationship between frequencies of precipitation and hot days [?]

Sorry. This should have stated the relationship between frequencies of precipitation and hot days.

7 Western Europe - which lat/lon region?

10W – 30E; 36– 61N: this will appear in the text.

10 tends to favor -> need some characterisation of goodness-of-fit

This is shown by the non parallel QR lines. This will be expanded on in the revised text.

13 high frequency of hot days is not the same as occurrence of heat waves; persistence is always part of the definition

Agreed. For clarity we will change to "hot summers" to avoid confusion. Hot summers in Europe yield a high probability of hot days [Parey et al., Clim. Dyn., 2000].

18 is this equating warmer conditions with dryer conditions? if so, this doesn't follow. if not, please clarify.

We will change the sentence to: "It was shown that the seasonal predictability of large European heatwaves decreases under drier conditions in southern Europe (10W -30 E; 36- 46N)), although their frequency increases (Quesada et al., 2012)."

25-6 confusing - you haven't looked at different climate forcings? or land use? horrible sentence...

Agreed. The point was to raise the issue that the connections between heatwaves and earlier precipitation might be affected by forcings (such as land use) over the millennium. The detection and attribution of such a change is work for the future and so we have rewritten this sentence for a clearer focus.

25-6 not clear at end of this section why bother to do any of it in terms of past for future. doesn't relate the two. doesn't avoid need for daily precip data. why not look at past and future simulations as well as present?

The point is that seasonal/monthly data are predictive of the relationships seen in daily data, and so the past1000 simulations can be used directly. Some past1000 runs will provide high frequency diagnostics as well though. We unfortunately have not yet completed these analyses.

p791 13 why select 5 (or is it a coincidence)? why not by sign of changes, or magnitude, or by cluster? Sign of change (quadrants of positive and negative dipole change) makes most physical sense to me, though I understand you may think the small magnitude changes (close to 0,0) are not conclusive.

We chose to divide the models into 3 groups, so for 16 models, each group must have 5 or 6 models. We tested the sensitivity to the number of groups, for example 2, and it doesn't change the results qualitatively.

16-19 "drier", "wetter" and "shift" are confusing - does this refer to present day (in which case clearer to say "drier Guyana...than the Nordeste..., associated with a southerly position of the ITCZ" etc), or a change between present day and future? or past? 20 It's

important to be incredibly precise when talking about anomalies of anomalies. Dipole changes are quite confusing and conflate two possible signals. If I've understood correctly, how about this: "Figure 4 shows that groups 1 and 3 have distinct precipitation responses to each other in both the past and future. In group 1 models, where Guyana is drier than the Nordeste in the present day, the change in dipole is negative for both past and future. This indicates that in both the MH and RCP8.5 simulations, either Guyana is drier and/or the Nordeste wetter than in the present day simulation. In group 3 models, where Guyana is wetter than the Nordeste in the present day, the dipole change is positive for both past and future, indicating that Guyana is wetter and/or the Nordeste drier than the present day."

No, the description of present-day precipitation in the suggested rewording is not correct. The present-day precipitation is similar for most groups, except for a tendency for group 3 models to exhibit more of a double ITCZ.

We will however clarify the wording in the revised version. Figure 4b shows simple anomalies and not "anomalies of anomalies", so to simplify the message we will focus on describing this figure. The "anomalies of anomalies" was just a metric to better represent the dipole structure and the past/future correlation.

This is why I don't understand the chosen grouping of models. Surely the sign change is more important than the size of change? Or are you saying that group 2 have too small a change to be confident?

In the revised version we will show 2 groups of models based on the sign of the dipole change, qualitatively it makes little difference.

23 broadening - does this contradict earlier "Northward shift" statement?
p792 13 atmosphere appears to play a key role -> Er...yes! Can this be replaced with something more useful?
19-26 Unclear generally.
20 Why link RCP projections and MH together, but 4xCO2 separately?

23 different precipitation response / similarity between -> contradictory?

We will clarify these issues in the reviewed submission.

p793 7 Best make it clear from the start this is a proof-of-concept estimate

Yes.

17 Schmidt et al. 2013 -> not in ref list

Added.

18 "may also differ" -> "differ" (and list which ones)

Added.

20 we expect -> list different c.s. values of models and model versions to support this

We will further quantify this.

p794 13 biases -> explain General: Explain in detail how Fig. 6 dashed lines and red dots are obtained.

This will be expanded

p795 1 somewhat sensitive - quantify

Yes.

10 I would think the mode was more useful than mean?

Both are useful - and have been included.

*11 what *is* the function - Gaussian?*

This will be clarified.

p796 21 looks like it starts in July not Aug?

Mostly true.

23 State that future anomaly is RCP8.5, 2036-2065; explain why this date range rather than e.g. end of century.

This is a period with a very large spread amongst models. Later in the century, almost all models show 'ice free' summers and so there is no discrimination between projections possible.

Is the Sep celestial or present calendar – and would it better to use the minimum rather than a fixed date anyway (to maximise signal)?

This is a reasonable point, but we used September in the present calendar in this analysis. We do not expect the results to vary greatly with this choice.

p797 3 What do the proxy data say?!

The paleo-data are described and referenced in the previous paragraph.

Is the conclusion that there is a "possibility to discriminate based on interannual modes of variability" as stated at the start of the section, or not?

This is a conclusion drawn in the cited paper, not a statement we are attempting to show here. Indeed, the paleo-constraints are independent of those claims.

6 connections -> not clear if this means spatial, temporal or both?

it is meant generally.

8 divergence -> explain / rephrase e.g. "changes in the climate proxy relationship" or similar
8 misunderstanding of -> poor understanding or representation of

This will be edited.

p797-799 Section 5.1. I find this section very confusingly written, and don't understand the point or conclusion. Is it this? "Palaeoclimate methods can be applied to the historical period too, if observations aren't available. There aren't many SST measurements

in the tropical Pacific to test climate models with, so we try a proxy sensitive to SST instead. First, we test a forward model of coral response by driving it with 1958-1990 observations of SST and SSS and comparing the output with that part of the coral record. Then we test climate models by using their 1890-1990 simulations of SST and SSS to drive the coral model and comparing with the longer coral record. The models and data disagree. We don't know if this is due to the climate model, the coral model, or the coral data. But there is no robust relationship between past and future anyway."

The reviewer's summary of this section is generally correct, but we will certainly work to revise the section for clarity. Although the climate models, the forward model and the coral observations may all play a role in the model-data discrepancy observed, work to date (and presented in the top panel of Figure 10) suggests that this discrepancy may be attributed (at least in part) to simulation of salinity. As discussed previously, these model-data comparisons are still useful in highlighting these potential biases in the models and data so that such metrics may be further developed and improved in the future.

In this paper, this relationship should be the first thing to be assessed. There is none, so this section is not necessary.

Demonstrating a lack of a relationship is useful. If one never mentions negative results, the literature ends up strongly biased and people waste time looking for results that others have already shown not to exist.

Even if the aim is to highlight the difficulty in diagnosing the reason for model-data disagreements when there is a modelling chain, it should be rewritten so that it is much more clear this is the point.

The point of the section wasn't to highlight the difficulty in diagnosing the reason for model-data disagreements, but rather to highlight how such comparisons may be used to identify potential biases in the models and proxy data that may be explicitly tested in current and ongoing work (e.g., see Thompson et al., 2013).

Thompson, D.M., T.R. Ault, M.N. Evans, J.E. Cole, J. Emile-Geay, and A. LeGrande, "Coral- CGCM comparison highlights role of salinity in long-term trends." P. Braconnot, C. Brierley, S.P. Harrison, L. von Gunten (eds) El Niño Southern Oscillation: observation and modeling, PAGES news 21(2) 2013 (in press)

p800-801 Section 5.2. Again, the point of this section is not clear at all. There is no reference to future simulations that I can see. Perhaps the point is to show that model-data disagreements are difficult to diagnose when there are substantial forcing uncertainties, but this is undermined by (a) using the wrong forcing for some simulations, therefore in principle overestimating the effect of uncertainties, and (b) providing no past-future relationship for, and therefore motivation to care about, the comparison.

Since we are unaware of what the 'right' forcing is, pointing out that some diagnostics are very sensitive to those uncertainties (even ignoring mis-specifications in some simulations) is still valid. We did not attempt to precisely quantify the effect for precisely those reasons.

p802-3 Section 5.3. Once more - this is not a past-for-future section. What is the conclusion: that PDSI is not useful because it cannot be calculated for models (and the reconstructions "would not be expected to line up". – does that mean matching specific wiggles of interannual variability..?), and that soil moisture is not useful because it has not been reconstructed (yet)? Sections 5.1, 5.2 and 5.3 should therefore be removed and put in a MIP paper that focuses on understanding models, not on constraining future projections. For this reason I haven't included specific comments for these sections.

People have already published projections of PDSI as if that was a comparable drought index to the PDSI reconstructed in, for instance, the N. American drought atlas (i.e. Dai, 2012). Whether paleo droughts have a forced component is still an open question. In the some models, the answer seems to be that N. American droughts are not obviously externally forced over the last millennium, even if similarly sized events occur in the

models over the whole period (e.g. Coats et al, 2013).

We will clarify why these sections are relevant in the revised manuscript.

*p804 3 For comparisons to be meaningful, well-characterised uncertainties should be stated as *essential* not recommended.*

Agreed.

7 I don't understand - why should the metrics not be sensitive to the model structure? (do you mean the definition rather than the value? if so can you give an example/section number).

One example from modern observations would be stratospheric circulation. This is heavily affected by the position of the model top and vertical resolution. An diagnostic that had a strong modulation by the QBO would not be a good candidate for constraining models that do not have the resolution or physics that give rise to the QBO in the first place.

7 I don't understand what "within the scope of the modelled system" means - if it means "should be based on simulated quantities" then this seems too obvious a statement to include...!

Well, sometimes this isn't so obvious. One example would be if a model does not include interactive dust, looking at dust deposition is not going to be a useful constraint since that variable is outside the scope of the modelled system. Oceans are outside the scope of an AGCM, sea level is outside the scope of a model without interactive ice sheets etc.

9 Who is assuming what relationships exist - can you give an example/section number?

This is more a reference to examples in the literature where casual claims that model skill in paleo-climates is important to future predictions are often made without this

actually being demonstrated. Note that this is why it matters that we report metrics where a relationship was AND wasn't found to be maximally useful for the community.

p805 8 point out that the difficulty with (in particular) the earlier eras is the absence of uncertainty assessment in the reconstructions

Yes.

9-10 high frequency diagnostics do not follow from more model experiments, and characterising structural uncertainty does not necessarily follow from high freq diagnostics.

These are of course separate developments and the wording will be changed to improve the clarity of the thought.

Table 2 does not seem to be referred to in the text.

This was an oversight.

Fig. 1 Perhaps it was thought too obvious to state, but I missed why the reconstructions are not shown for the other two regions – not a difficulty in combining land and ocean, if they are only over reconstructed grid boxes?

Indeed. The upper two panels show averages over the whole region - not just where there are reconstructions.

Fig. 3 Contradiction: Southern or Western Europe? Is it different for temp and precip? "Consistency" -> insert "qualitative", unless a more specific test than visual inspection has been done? Better to have narrower regression lines for clarity, and ideally label them to be really clear.

Precipitation frequency is computed over Southern Europe, between Jan. and May. The hot day frequency is computed for Western Europe between June and August. The figure caption will be changed. The two QR lines will have different colors for clarity.

Fig. 4 Legend should switch Group 1 3 labels. For clarity, I think zero lines should be marked in (a), and the regional boxes marked on all maps (though without labels for the other 3). Ideally the white areas would be low signal-to- noise – for example, using a rank sum test – rather than a fixed threshold.

The figure will be redrawn to account for these comments and those above.

Fig. 5 Why no model labels this time? Why not white for small anomalies on the maps again? It might be an artifice of the pattern, but (b) and (c) look buggy – why the strong horizontal divide in the south – is it sea ice? Worth a comment. Caption for (b) needs to say LGM in it.

Agreed.

Fig. 6 Why no model labels? Linear correlation -> linear regression? Predictive uncertainty -> bootstrapped confidence intervals? Observational value -> reconstruction?

Replaced in the revised text.

Fig. 7 Show underlying histogram to indicate how well kde fits data. Obviously sensitive to bin width, but would show clearly what is being done and why it is so sensitive to the number of ensemble members.

This will be added.

Fig. 8 Historical -> early historical, or preindustrial. Draw zero line in (c) to help make the point in the text. Ideally, use dashed/dotted lines to distinguish between similar colours (e.g. pinks/oranges).

Yes.

Fig. 9 Caption – refer to Fig. 8 b and c. Use different symbols for models.

will do.

Fig. 10-13 No detailed comments – see text.

Captions and figures will be clarified along the same lines as in the text.

Minor edits

These have all been incorporated in the revised text where still relevant.

Interactive comment on Clim. Past Discuss., 9, 775, 2013.