

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.

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

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. 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.

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?

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. 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. 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.

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. 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).

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).

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).

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. See list of specific comments for some more details.

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;
- 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.

Below are specific comments, followed by minor edits that may be useful if the authors revise the manuscript.

Specific comments

p781

14 clarify physically meaningful

21 there are other types of forward modelling. e.g. within a proxy model with a climate prior that is not from a climate model - "or other climate prior" would cover it

p782

12 add later ref

19 clarify "statistical uncertainty" -> alternative term for measurement uncertainty? (if so, delete) or something else? (if so, clarify)

26 data-model comparisons -> if these cannot be expressed with Euclidean distance.

p783

8 relatively large forcings -> clarify - larger than millennium but smaller than LGM

10 forcings are much smaller - clarify that not including LGM and MH

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

27 and that agreement may be due to error cancellation.

p787

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

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

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.

p788

9 Sutton et al. not in ref list.

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...

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?

p790

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

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

7 Western Europe - which lat/lon region?

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

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

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

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

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?

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.

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."

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?

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?

p793

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

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

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

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

p794

13 biases -> explain

General: Explain in detail how Fig. 6 dashed lines and red dots are obtained.

p795

1 somewhat sensitive - quantify

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

11 what *is* the function - Gaussian?

p796

21 looks like it starts in July not Aug?

23 State that future anomaly is RCP8.5, 2036-2065; explain why this date range rather than e.g. end of century. Is the Sep celestial or present calendar - and would it better to use the minimum rather than a fixed date anyway (to maximise signal)?

p797

3 What do the proxy data say?! 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?

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

8 divergence -> explain / rephrase e.g. "changes in the climate proxy relationship" or similar

8 misunderstanding of -> poor understanding or representation of

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."

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

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.

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.

p804

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

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).

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...!

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

p805

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

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

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

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?

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.

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.

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.

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

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.

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).

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

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

Minor edits

p778

1 20th Century (starting from 1850) -> from 1850 to 2000

3 30 yr -> prefer “years” (throughout)

p779

8 to situ -> to in situ

p780

5 roughly produce -> correspond to?

6 have been -> that have been

23 periods when are -> periods are

p782

9 correlations to -> correlations with

25 take account data -> take account of data

p783

20 For example, both the magnitude of solar or volcanic forcing over the last millennium, and the size and height of ice sheets at the LGM are -> For example, the magnitudes of solar and volcanic forcings over the last millennium, and the size and height of ice sheets at the LGM, are sources

27 was related -> were related

p784

7 and weighting -> and (any) weighting

16 fields of modern data and model output have been commonly used in model evaluation -> fields of data and model output have been commonly used in model evaluation for the modern observations.

p785

2 Recent work, has -> Recent work has

3 timescale, and in principle, these -> timescale and, in principle, these

p786

2 correlation of model output -> correlation between past and future in the model output

6 physics ensembles is more explicitly Bayesian, considers -> physics ensembles, is more explicitly Bayesian and considers

11 prior constraint, which depending on one's viewpoint, and the specific ->
prior constraint which, depending on one's view point and the specific
12 of this approach -> delete

p787

24 surprisingly -> surprisingly (considering the proximity to the ice sheets)

p788

6 land -> land (excluding ice)

7 “, a result close to ... future climate.” -> “, and ~1.5 in both the PMIP2 fully
coupled LGM (Braconnot et al., 2012) and CMIP3 future experiments (Sutton et
al., 2007).”

12 well-documented - by who? w.r.t. models or data? why does it matter to this
sentence?

p791

12 onwards Simplify lat/lon regions: e.g. 5degS-15degS-45degW-35degW -> 45-
35degW

27 missing full-stop

p792

6 Table 1 -> 2

12 New paragraph at Second.

14 amipFuture simulations -> the atmospheric model intercomparison (AMIP)
future simulations ("amipFuture"), where

16 already -> delete

17 fast response -> fast atmospheric response

21 They also happen to be the models with -> These models also show

26 Finally -> new paragraph

27 it is worth nothing that -> delete

p793

7 estimate of -> estimate

18 equivalent -> delete

24 biases, -> biases:

p794

3 anomaly -> delete

16 New paragraph for referring to Fig. 6.

21 Remove new paragraph, or refer to Fig. 6 again.

22 observational estimate -> reconstruction

28 make clear if this is a quote from paper or new numbers

p795

1 MC sampling -> bootstrapping?

12 maybe -> may be

13 variability -> of volume?

14 a orthogonal -> an orthogonal

p796

20 sea ice extent -> add "relative to the present day" otherwise could be ambiguous (could mean intra-year)

21 winter months -> add "(Dec-Feb)"

21 are not coherent -> do not agree in sign

27 Although, -> delete comma

p797

9-11 Rather clunky and unclear.

p802

7 precipitation sensitive -> insert hyphen

9 temperature-derived -> insert hyphen

15 how well does this index ... reflect -> how well this index ... reflects

23 Beginning near the middle of ... begin -> In the middle of ... begin

26 actually -> delete

p803

8 forcing -> changes

14 focused -> focused on

14 by inclusion of -> delete

18 the ensemble of models, simulations and paleo-data -> models and paleo-data

OR models, paleo-data and climate states?

20 correlated to -> correlated with

p804

5 should -> should

10 significant -> reliable

16 should -> will