Clim. Past Discuss., 8, C2372–C2396, 2012
www.clim-past-discuss.net/8/C2372/2012/
© Author(s) 2012. This work is distributed under
the Creative Commons Attribute 3.0 License.

Climate of the Past Discussions

Interactive comment on "Climate of the last millennium: ensemble consistency of simulations and reconstructions" *by* O. Bothe et al.

O. Bothe et al.

ol.bothe@gmail.com

Received and published: 16 November 2012

Dear editor, dear referees,

In the following we present our final response to the three referees' comments. Please see attached diff-file for all changes. We are going to refer to this file throughout our reply. Please consider also our initial replies since we bascially followed the work that we have outlined there.

First we want to repeat our acknowledgement of the referees' contributions which

clearly helped to improve the quality of our manuscript.

We will comment on each of the reviewer's comments and afterwards add a description of how the manuscript changed. To some extent, the present reply is going to quote the initial replies or just refer to them.

Referee #1

Referee #1: page C1363: *This is an interesting idea and analysis, but some aspects of it are rather unclear to me.*

In general, throughout the manuscript, I often found it hard to follow exactly what was being tested against what. It would be very helpful if some clear terminology could be set out and consistently applied. It seems that the ensemble mean (of model simulations, or proxy reconstructions) is being tested for compatibility as a sample from the ensemble of reconstructions or model simulations respectively. Some of these analyses appear to have been done both ways round, but others (figs 4 onwards) only in one direction. It would be useful to have a clearer statement of the hypotheses being tested (eg, that a specific validation target was drawn from the same distribution as a particular ensemble).

We hope that our changes from the abstract to the description of the results clarify sufficiently, the tests, the terminology and the hypotheses. Please refer to pages 1–10 of the diff-file for all made changes.

Referee #1: page C1364: *The authors state in their abstract that "no status of truth can be assumed for climate evolutions...". But what is meant by a status of truth*

anyway? I'm afraid I don't grasp the point being made.

We changed the abstract to clarify this point. "Status of truth" referred to an accurate description of an unknown past climate state. Our conclusion is that such a description cannot be inferred from the notably differing reconstructions and simulations. Note, although we state the differences, we assume reconstructions and simulations to be equitable hypotheses about the past climate.

Referee #1: page C1364: *I think the authors also need to be more careful with their mention of a "true" distribution eg on line 25 of p 2411. What exactly is "truth" here? I think a more accurate description of the situation is that the historical climate system had a true state (or vector of truth through time), and any uncertainty, which may be represented through a distribution, is solely due to our limited knowledge. The concept of truth therefore is not applicable to the distribution itself. More generally, I suggest looking for instances of "truth" or "true" in the manuscript (which is usually presented in quotation marks, already suggesting some discomfort with the term) and considering whether the concept at hand can be better expressed. In the case of a verification target (eg Fig 3), it is probably uncontroversial, but as a description of a distribution, it seems inappropriate.*

Beyond our initial reply we continuously changed references to "truth" and "true" data to references to "target" or "verification" data. With respect to the distributions, we think that all discussions of them are now appropriately formulated. We further added clarifications to the Abstract, the Introduction and the Methods-section. Please see the diff-file for the specific changes.

Referee #1: page C1364: *The use of ensemble means as verification target introduces a confounding factor due to smoothing of internal variability. This is briefly mentioned in the manuscript, but does not appear to be adequately treated. In the*

simple perfect case where all time series for both ensembles are drawn from the same generating method, the ensemble means will tend to show inadequate variability in a naive comparison. Perhaps this underlies the result summarised as: p2421 l13-15 "Figure 1 further shows that the considered ensembles of estimated temperature anomaly series generally enclose the verification data (Fig. 1a–c), but they often over- estimate inter-annual variability (Fig. 1d–f)." This issue should be addressed more clearly, eg by augmenting the ensemble mean verification target with internal variability "noise" (contrast with the use of noise on the ensemble members, to account for observational errors on the target).

We have repeated the analyses based on additional estimates of internal variability. See the initial replies and please refer to the diff-file for changes to the Methods- and Results-sections. This also is reflected by changes in the Discussion.

The re-calculation changed the results. Inconsistencies are generally less pronounced or may even vanish. Additional sensitivity discussions clarify that our conclusions are still valid.

Referee #1: page C1364: *Incidentally, the Hind et al manuscript is now published in two parts, as the authors may have already noticed.*

Thank you. We changed the references to Hind et al. (2012) and Sundberg et al. (2012) as appropriate.

Referee #2

Referee #2: page C1508: *The authors admirably approach an increasingly important area of research involving the comparison of both the simulated "model" climate and that reconstructed from proxy information, however the article has a few shortcomings regarding the layout, structure and language which make it, at times, difficult to follow. I have attempted to discuss my desire for the clarification of a few things below in more detail though I want it noted that I find the analysis technically sound and commendable for stepping away from, as stated by the authors themselves, the "wiggle matching" or "by eye" approaches traditionally used in comparing simulated and reconstructed data.*

We thank the referee for his/her appreciation of our work.

Referee #2: page C1508: *I often found myself confused by what was meant by consistency, though I am aware that these terms are comprehensively dealt with elsewhere. Although a detailed discussion occurs in section 2.1 to inform the reader, you also discuss "reliability of a probabilistic ensemble" which I also found confusing. For example, the authors could reference more clearly that this sentence stems from forecast analysis and is now a methodology being applied to climate science (if I have understood this correctly).*

We hope that our continuous changes to the text clarify the concept, its origin and the terminology. Please see especially pages 1–10 of the attached diff-file.

Referee #2: page C1509: *I would like help with understanding how "deviations in climatological distributions" (page 2414, line 28) can be interpreted in the reconstructed and simulated worlds. I would also welcome a more thorough discussion of what the probabilistic and climatological components of consistency are in the*

analysis. Perhaps a brief clarification of the similarities and differences between the rank histogram analysis and the r-q-q plots (from Marzban et al. 2010) would also link well with these clarifications, where you could translate these forecast ideas and methodologies into this climatological comparison framework to help the reader understand why you use both methods here. This discussion could be where we find the discussion to help interpretation of the rank histograms (page 2413, line 22).

Again, we hope our changes in the Introduction and in the Methods-section improve the understanding of our argumentation. This holds especially for the discussion of the two components of consistency and their interpretation in the climate context.

Referee #2: page C1509: *I have seen the previous referee's comments and the author's reply, however I would reiterate that whilst your definition of "truth" or "true" is discussed somewhat and referenced towards Annan and Hargreaves (2010) for example, there are instances where things are simply not clear such as page 2413, line 23: "If the truth is sampling from a distribution narrower (wider) than the ensemble". I realize these ideas and definitions have already been represented in previous research but is it not possible somehow to distinguish between the "observable truth" and the "true" mean or distribution of a data set by using a more distinct and explicit terminology? i.e. simulated ensemble mean target / reconstructed ensemble mean target?*

We followed the common suggestion by all Referees and changed the manuscript accordingly. We hope that our modifications to the description of the applied methods make them clear enough to understand our argumentation and our approach.

Referee #2: page C1509: *Regarding section 2.2, it is not clear to me why the approach is reversed only for the Frank et al, (2010) data and not in the Central European and global temperature field consistency analyses. Note here, I find the in-*

clusion of the "single forcing simulations as a valid hypotheses about the pre-industrial climate trajectory" page 2415, line 11, as likely to be useful in broadening uncertainty estimates regarding pre-industrial forced components.

The mutual assessment depends on the availability of ensembles of both reconstructions and simulations. The only available reconstruction ensemble is the data by Frank et al. (2010) for the hemispheric mean temperature. Production of an ensemble for a seasonal central European temperature would be a project and a publication on its own. We did not pursue this task mainly because of the arbitrariness in defining a common seasonal and areal Central European coverage and the amount of common input to some of the available data sets. Constructing an ensemble of global or hemispheric field reconstructions is clearly beyond the scope of the present study.

Referee #2: page C1510: *In regards to the order of some results discussed in the main text, there is a good results summary at the start of section 3.1.2, but I wondered if it might be better to discuss Fig. 1 earlier in the text? I felt the same for Fig. 4c which is discussed in great detail in section 3.2.2 after later figures are expanded upon earlier on in the results discussions.*

We extended on the discussions of old Fig. 1 and Fig. 4 (now Fig. 3–5 and Fig. 9) at the beginning of the respective Subsections. We hope the respective changes increase the readability of our manuscript.

Referee #2: page C1510: *Page 2423: Here you refer to the "first" or "last" records, forgive me if I have missed something but I don't understand what these are or how they are defined. You could expand upon line 1-3 describing the different periods. You also mention "early" and "late" records, how do these fit in with the first, second, third and last described in the caption of Fig. 5? I think it would help if you expanded upon your description of the Fig. 6 results generally here as well. I.e. page 2424, line 20:*

Here you refer Fig. 7e as being "about 1595 to 1845". This specific terminology is helpful and could be applied earlier in your results.

We clarified the use of the sub-periods in the Methods-section and at the start of the Section on the spatial field data.

Referee #2: page C1510: *Page 2424, line 25-29: These statements feature significant conclusions, perhaps the motivation for these statements should be shown in Figures? or at least in supplementary material.*

These statements changed due to changes in the presentation of the results in reaction to the review by Dr. Edwards. We nevertheless add a panel to the former Figure 4 (now Figure 9).

Referee #2: page C1510: *Page 2424, line 10: What is meant by a "moderate random error"?*

We clarify the choice of uncertainties at that position, in the Methods-section and where further thought appropriate.

Referee #2: page C1510: *Page 2425, line 6: Could you expand upon what "an error estimate" is defined as here?*

Again, we clarify the choice of uncertainties at that position, in the Methods-section and where further thought appropriate.

Referee #2: page C1510: *Section 4: This introduces a lot of new results into the analysis which are in addition to those discussed earlier. It is good the authors contextualize their work here but there is a great deal of new analysis introduced*

here that is not shown (perhaps it could be included in supplementary material?). Is there some way you can integrate these additional results more clearly with the earlier analysis so the reader can see how and where in your analysis these additional results (and the work by other authors) can be compare?

Most of the additional results are now described in the Results-section and discussed in this context. We add additional Figures. Please see diff-file.

Referee #2: page C1510/1511: *Some further comments about the otherwise excellent Figures: Figure 2: You have results for both an ensemble mean simulation Fig. 2d for one area averaged NH series, and for the ensemble mean reconstruction equivalent Fig. 2f next to one another. What are the implications for the results in using these different approaches? I was also unclear what "sub-ensemble" refers to in regard to the results of Fig. 2e and 2f.*

We discuss the aspects and the implications of the different approaches in the Methods-section and in the Results-section. Please refer to the changes highlighted in the diff-file.

Referee #2: page C1511: *Figure 4: I would appreciate further discussion regarding the "mapped ranks" and how they were derived.*

We extend the description of the mapped ranks given in the Methods-section.

Referee #2: page C1511: *Figure 5: I think it would help if you highlight in the caption of Fig. 5 that when you say "left to right for the first, second.." you are discussing the coloured box legend in each individual histogram.*

We followed the suggestion.

Referee #2: page C1511: *Incidental remarks: Page 2418, line 16: "Contrarily," could be phrased "In contrast,"*
Page 2418, line 23: add an "a" here: "but originates from only a few ensemble"
Page 2420, line 10: becomes "simulation ensembles and reconstructions."
Page 2423, line 20: you can remove the "n" from "neither" and from "nor".

We considered the incidental remarks.

Referee #3: Dr. T.L. Edwards

Dr. Edwards, referee #3: page C1800: *This manuscript is quite poorly-written and hard to understand. There are several aspects in which the authors do not convince me they have understood, or at least sufficiently addressed, the scientific and statistical pitfalls of applying rank histograms (and related analyses) to palaeoclimate data.*

We hope our changes account for Dr. Edwards substantial criticisms. See the changes highlighted in the diff-file from the Abstract to the Results-section.

Dr. Edwards, referee #3: page C1800: *These comments apply to the manuscript up to and including Section 3.1.1.*

We honestly regret that Dr. Edwards didn't find the time to comment on the further parts of the manuscript since her comments notably improved the paper.

Dr. Edwards, referee #3: page C1800: *General scientific issues a) Incorrect scientific questions / inferences - The abstract states that the simulated and recon-*

structed ensembles are both over-dispersed. I think this is a consequence of using an ensemble mean as a target (lower interannual variability). I don't think ensemble mean targets, and therefore these conclusions, are valid.

As mentioned above, we re-calculated much of our analyses considering the reduced variability of the ensemble mean targets. We hope that these major changes account for the concerns of all three reviewers.

For completeness sake we also repeat our initial reply: "...The current version of our revised manuscript addresses this concern by including estimates of internal variability. ...Indeed the over-dispersive character is not as pronounced as before but is still the most common inconsistency since the abstract also refers to the field reconstruction...

However, we agree that it is necessary to (i) justify the use of the ensemble mean and (ii) to discuss here part of the results on page 2418.

In the original manuscript we assume that uncertainties for an ensemble mean which are proportional to the spread of the respective ensemble. In principle, these uncertainties account for the differences between the assumed signal and how the individual ensemble members represent it. These differences can be assumed to represent a combination of the internal variability and the methodological uncertainties. If we randomly sample these uncertainties to inflate members of the ensemble that we wish to verify, this should compensate for the reduced internal, i.e. unforced, variability in the ensemble mean under certain conditions. The main condition is that the forced signals represented by the ensemble mean target and the individual ensemble members are similar. It further implies that these uncertainties are uncorrelated in time. In our understanding the analysis remains valid considering our original discussion of the caveats in sections 2.3 and 4. Obviously the analysis benefits if we address these issues directly by adding estimates of the internal variability to the verification target as we do in reply to referee 1."

The re-calculations force us to modify our conclusions with respect to parts of our analyses. However, the general conclusions of the manuscript are not changed. Please see all changes made to all sections in the attached diff-file.

Dr. Edwards, referee #3: page C1800: *- Incorrect inferences from figures - see specific comments for p2418 below*

We reply below the specific comments.

Dr. Edwards, referee #3: page C1801: *b) No discussion / addressing of correlation of reconstruction ensemble members*
- Correlation across reconstructions is mentioned by Frank et al., and described as a problem by Marzban et al., but not mentioned or addressed
c) No discussion / addressing of correlation of simulation ensemble members
- As above. I'm not convinced that a forcing ensemble is a valid subject for this kind of analysis. If it is, it requires thoughtful discussion, and addressing of, the high degree of correlation between ensemble members due to the design of the ensemble (limited number of possible forcing inputs).

We discussed the not-so-high degree of correlation in our initial reply. In contrast to our initial reply, we now regard this problem as of rather reduced importance and therefore discuss it only shortly in the revised draft. Please refer to the diff-file. For completeness sake, we repeat here part of the initial reply:

"Next, we will first respond to comment (c). We note that for the European data, the maximum inter-ensemble correlation is about 0.17 and it is usually between 0.2 and 0.25 for the northern hemisphere data. The former is notably stronger than the correlation between target and individual simulations. The latter is usually in the range of correlations between simulation members and the ensemble mean reconstruction target.

We quote here the following from Marzban et al. (2011): '...if the correlation between ensemble members and the observation is generally different from that between ensemble members, then the RH cannot correctly assess the climatological component of reliability. For realistic ensembles, where both correlations are large (õ0.9), the RH is then expected to be U shaped by default, and so cannot assess the climatological component of forecast reliability. And if the ensemble forecasts are more similar to each other than to observation, then the RHs are still U shaped, even though there exists no climatological overdispersion.' Please note that from our understanding the end of the last sentence should read "underdispersion" instead of overdispersion. Please, refer further to Figures 5 and 6 of Marzban et al. (2011). For all possible readers of this reply we note that an author version of Marzban et al. (2011) is available at Caren Marzban's homepage at the University of Washington (http://faculty.washington.edu/marzban/). From the given numbers, we would assume that our assessment should reasonably approach the uniform outcome and should not show deviations as large as found in our evaluation. ...

Comment (b): The intra-ensemble correlations for the reconstructions (see below for definition) is usually between 0.4 and 0.6, but reaches 0.82. The correlation to the ensemble mean simulation target is usually smaller.

Thus we would expect under-dispersive u-shaped rank-counts following Marzban et al. (2011) for all three area-averaged assessments."

Dr. Edwards, referee #3: page C1801: *d) Insufficient discussion / sensitivity studies of effects of temporal correlation*
- How were the bootstrap lengths chosen, and what is the sensitivity to this choice?
- How is the temporal correlation estimated (degrees of freedom / correlation coefficient) for the chiËĘ2 test?
- What is the sensitivity of the result to this choice?

While we agree that a discussion of the sensitivity to made choices is important, we do not regard it as such a major problem. Therefore our related changes are rather short in the Methods-section and mainly follow our initial reply to Dr. Edwards parts of which we repeat here:

"The bootstrap block lengths: We subjectively applied a 50 year window. Generally it can be noted that longer windows reduce the width of the quantile ranges while shorter blocks widen it. The 50 year block length primarily was a compromise between the preferable lengths (according to the auto-correlation functions) for the ensemble mean target (longer window) and the simulation ensemble members (where shorter blocks are possible). Thus, for the case of the ensemble mean target we indeed use possibly too short blocks.

The degrees of freedom of the time-series are estimated following [the literature, compare e.g. Bretherton et al. (1999). This] enters the calculation of the χ^2 statistic directly and automatically.

As discussed in the referenced literature and also mentioned in the draft, this has an essential influence on the results."

As we changed the presentation of the results following Dr. Edwards' suggestions, this is not any longer easily seen from the results. We hope that the informations in the Methods-section are enough to highlight these sensitivity.

Dr. Edwards, referee #3: page C1801: *Other manuscript issues*
2411/18 - There is no "status of truth" even when there are observations: our knowledge is always imperfect.

We changed this.

Dr. Edwards, referee #3: page C1801: *2411/20, 2413/8 - This is not an "objective" measure because arbitrary choices are (necessarily) made throughout.*

We modified the use of "objective".

Dr. Edwards, referee #3: page C1801: *2414/18 - Explain bootstrapping method and motivation more clearly, for those not familiar with it*

As already stated in the initial reply we prefer to rely on a reference to Efron and Tibshirani (1994).

Dr. Edwards, referee #3: page C1802: *2414/24 - Explain climatological and ensemble components of reliability*

We hope our changes to the Methods-section account for this comment.

Dr. Edwards, referee #3: page C1802: *2415/11 - Add table of simulation experiments to clarify forcing combinations*
2415/11 - Add figure showing forcing timeseries
2425/11 - Explain how weak and strong solar forcings are chosen - bounds of an uncertainty range?

With respect to all three comments, we want to note, that we do not think table, figure and discussion would improve the information content of the manuscript. Indeed, we think it would reduce the readability. Please note that Jungclaus et al. (2010) is freely available and gives the relevant information.

Furthermore, our analysis and our conclusions do not depend on the explicit structure of the ensemble but simply on the use of an ensemble.

Dr. Edwards, referee #3: page C1802: *2415/17 - Are all reconstructions annual? (Sentence implies not all are)*

We hope our modifications of the sub-section on the data clarify this.

Dr. Edwards, referee #3: page C1802: *2415/18 - What are the different members of the reconstruction ensemble? How many are there?*

We hope our changes throughout the manuscript clarify this.

Dr. Edwards, referee #3: page C1802: *2415/22 - What are the sub-ensemble members, how many are there, and how were they chosen?*

We hope our changes to the section on the data clarify this.

Dr. Edwards, referee #3: page C1802: *2416/12 - "similar caveats" - which? What is their implication here?*

We have clarified this.

Dr. Edwards, referee #3: page C1802: *2416/23 - Needs better explanation of how adding error to the target can compensate for temporal smoothing*

We do not extend on this discussion since we changed the analysis to targets with added internal variability estimates. This is discussed in the revised manuscript, please see diff-file. We hope this is sufficient.

Dr. Edwards, referee #3: page C1802: *2417/15 - Comment on / explain Fig. 1 here, especially righthand column*

We hope our changed discussion of Fig. 1 accounts for this comment.

Dr. Edwards, referee #3: page C1802: *2417/21 - Remove all plots without observational errors added - this is not a study of the method but an application*

Done

Dr. Edwards, referee #3: page C1802: *2418/14 - Explain this result explicitly: the simulated ensemble mean is generally near the centre of the ensemble of reconstructions, while the reconstruction ensemble mean is generally near the centre of the ensemble of simulations. Comment on why this is - presumably because the target is an ensemble mean and therefore has lower interannual variability. Is this a valid target? I don't think so (see Scientific Issues above). How about repeating this using one (or each) of the individual ensemble members as a target?*

We do now account for the reduced variability. We shortly discuss the consistency with respect to the individual sub-ensemble members.

For completeness sake, we add our initial reply to this comment: "the lower panels of Figure 1 [now Figure 4a and 5a] are indeed an expression of the use of the ensemble targets, but similarly reduced variability occurs at least for certain periods for a number of temperature reconstructions that are meant to reliably represent interannual variability (not shown). Furthermore, we discussed in the original manuscript: 'it is arguable whether an ensemble mean represents unfiltered annually resolved data. A posteriori, our approach seems to be valid for the comparison of the specific simulation ensemble mean with this particular reconstruction ensemble, but the larger variability in the simulations compromises the inverse consideration.' This note in part referred to [old] Figure 1, where we see that the proposed centered character of the simulation ensemble mean target relative to the reconstruction ensemble is much less pronounced

than for the reconstruction ensemble mean target relative to the simulation ensemble. [Old] Figure 1c rather indicates various periods where the ensemble is likely biased compared to the target.

Thus, while agreeing with the notes by referee 1 and 3 an ensemble mean is a priori a valid target. Following Persson (2011), Hargreaves et al. (2011), Johnson and Bowler (2009) and Marzban et al. (2011) we could have inferred a priori that the analysis would result in dome-shaped rank histograms and positively sloped residual quantiles. However, this is not true for the assessment of the bootstrapped intervals.

Less easily a priori inferable are the results obtained from the data if we add an additional estimate of internal variability. This analysis is included in the so far revised version of the manuscript. There we still see over-dispersion for the simulation ensemble even though it is reduced. However, the reconstruction ensemble is found to be probabilistically consistent, but diverse climatological deviations still remain. The current status of the manuscript follows the comment of referee 1 and discusses this more explicitly."

Dr. Edwards, referee #3: page C1802: *2418/17 - Give the critical values for the chiËĘ2 test*

We discuss the critical values in the modified Methods-section.

Dr. Edwards, referee #3: page C1803: *2418/17 - "highlights the problem of..."strictness" - I would say that it actually indicates a problem: for example, using the wrong number of degrees of freedom.*

We do not think so but hope that our short comment on the degrees of freedom in the Methods-section accounts for this comment.

Dr. Edwards, referee #3: page C1803: *2418/20 - "as shown by..." - I don't understand how, when the bootstrapped envelope encompasses the target zero line.*

We hope our changes clarify this.

Dr. Edwards, referee #3: page C1803: *2418/21 - "as shown by..." - explain how - is this because the grey band is slightly lower than the target? I can't see how this is worse in the 16th century.*

Due to changes in the description of the results this formulation is obsolete.

Dr. Edwards, referee #3: page C1803: *2418/24 - The slope is within the bootstrapped intervals, therefore should not be considered a slope*

We hope our changes clarify this.

Dr. Edwards, referee #3: page C1803: *2418/28 - I think this should be 3c,e not 3b,c*

Changed

Dr. Edwards, referee #3: page C1803: *2419/11 - Does this chiËĘ2 test include autocorrelation? If not, why use the bootstrapped intervals if these are then ignored in favour of a chiËĘ2 test under an assumption of independence?*

Yes, it accounts for autocorrelations, see the modified Methods-section.

Dr. Edwards, referee #3: page C1803: *2419/14-17 - "not shown" vs "compare Fig. 3" This section is not clear as to what is shown and what is not. Rank histogram*

not shown? Or weak forcing ensemble not shown? Shown in Fig. 3c-f or only c+d?

Since the representation of results changed the discussion of the display changed as well.

Dr. Edwards, referee #3: page C1803: *2419/19 - To improve clarity move "not shown" to after "truth"*

We made more substantial changes.

Dr. Edwards, referee #3: page C1803: *2419/19 - "into perspective" - what does this mean: confirms?*

It meant that the climatological assessment showed that despite the probabilistic consistency the climatological quantiles lack consistency.

We made more substantial changes.

Dr. Edwards, referee #3: page C1803: *2419/23 - bootstrap "generally" and "otherwise" - be more specific: a + e vs a only?*

We made more substantial changes to the text.

Dr. Edwards, referee #3: page C1803: *2419/26 - "amplified picture" - what does this mean? It seems to contradict line 19.*

We hope our changes clarify this.

Dr. Edwards, referee #3: page C1803: *2420/5 - "strong deviations" - presumably this is in R-Q-Q plot: be more specific, otherwise it sounds as though it is a*

deviation in the signal (e.g. bias) rather than quantiles (dispersion). I don't understand the rest of this sentence.

We hope our changes clarify this.

Dr. Edwards, referee #3: page C1803: *2420/8 - 50yr moving average has not been mentioned or explained before this point. I don't understand the rest of this sentence.*

Was removed.

Dr. Edwards, referee #3: page C1804: *Figure 1 - Too small and very unclear. Split into 3 (or more). Add individual legends to each to clarify confusing colour schemes. Remove ensemble means for clarity. Why is the moving average only shown for CE? This should be full size. I think there are both solid and dashed black lines but this is not explained. Why is the solid one so thick?*

We changed the Figure and split it in three.

Dr. Edwards, referee #3: page C1804: *Figure 2 - Remove all values of chiËĘ2 where autocorrelation is not accounted for - this is not a study of method*

Done

Dr. Edwards, referee #3: page C1804: *Figure 3 - Some of these colours look the same to me*

We hope that our changes to the Figures account for this.

Dr. Edwards, referee #3: page C1804: *Throughout:*
- I find the parenthetic structure for two-part sentences hard to read; suggest rewriting in separate sentences

We tried to account for this criticism. Please refer to diff-file for all changes made in the manuscript.

Dr. Edwards, referee #3: page C1804: *- replace "truth" with "target"*

Done

Dr. Edwards, referee #3: page C1804: *- update Marzban et al. 2010 ref to 2011*

Done

Comments on diff-file

Abstract

Changes in the Abstract account for the referees' comments on the terminology (e.g., "truth"), the modified calculations, and the comment on our concluding paragraph by referee #1.

Introduction

Various comments led us to clarify the language and content of the Introduction. We hope that now our approach and our argumentation can be more readily understood.

Methods and data

All referees commented on a lack of clarity in the description of the methods and our argumentation. Further notes implicitly or explicitly asked for a re-calculation of our analyses under different assumptions. We think that the extension of the Methods- and Data-sub-sections as well as of our discussion of the approach sufficiently consider the multiple annotations. Some short notes are new and deal with the sensitivity of the results to made choices.

Results

We add an initial subsection on the Intra-ensemble consistency as reaction to the comment on the additional results previously presented in the Discussion. This includes two new Figures.

We split the old Figure 1 into three Figures and discuss them in the beginning of subsection 3.2.

The comments on the use of an ensemble mean target and on the presentation of certain results led to changes in the subsequent Figures and descriptions. Note that the results change (see especially new Fig. 6 and 7). We further omit references to "truth" and change references to "errors" to "uncertainties".

We extend on the discussion of the results of the simulation ensembles relative to individual reconstructions. Therefore we add an additional figure (new Figure 8).

We add a discussion of the implications of the chosen approach for the results.

Our discussion on the origin of inconsistencies changes due the re-calculations.

Comments on the former Figures 4 and 5 and on former Section 3.2 (now Section 3.3) were considered in the modifications of the text. This includes results previously not shown, discussions of former Figure 4 (now Figure 9), clarifications on the sub-periods and more detailed captions.

We further follow the note that we should generally rewrite part of the manuscript due to the used structures of sentences.

Discussion

The Discussion changes since we moved part of it into the Results-section (e.g. new Section 3.1) and since the re-calculations forced us to reconsider some conclusions.

Concluding remarks

The Concluding remarks change since the re-calculations forced us to reconsider some conclusions.

References

Annan, J. D., Hargreaves, J. C., and Tachiiri, K.: On the Observational Assessment of Climate Model Performance, Geophysical Research Letters, 38, L24 702, doi:10.1029/2011GL049812, 2011.

Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M. and Bladé,I.: The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field, Journal of Climate, 12, 1990–2009, 1999.

Efron, B. and Tibshirani, R. J.: An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), Chapman and Hall/CRC, 1 edn., http://www.worldcat.org/isbn/0412042312, 1994.

Frank, D. C., Esper, J., Raible, C. C., Büntgen, U., Trouet, V., Stocker, B., and Joos, F.: Ensemble reconstruction constraints on the global carbon cycle sensitivity to climate, Nature, 463, 527–530, doi:10.1038/nature08769, 2010.

Hind, A., Moberg, A., and Sundberg, R.: Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 2: A pseudo-proxy study addressing the amplitude of solar forcing, Climate of the Past, 8, 1355–1365, doi:10.5194/cp-8-1355-2012, http://www.clim-past.net/8/1355/2012/, 2012.

Jungclaus, J. H., Lorenz, S. J., Timmreck, C., Reick, C. H., Brovkin, V., Six, K., Segschneider, J., Giorgetta, M. A., Crowley, T. J., Pongratz, J., Krivova, N. A., Vieira, L. E., Solanki, S. K., Klocke, D., Botzet, M., Esch, M., Gayler, V., Haak, H., Raddatz, T. J., Roeckner, E., Schnur, R., Widmann, H., Claussen, M., Stevens, B., and Marotzke, J.: Climate and carbon-cycle variability over the last millennium, Climate of the Past, 6, 723–737, doi:10.5194/cp-6-723-2010, 2010.

Marzban, C., Wang, R., Kong, F., and Leyton, S.: On the Effect of Correlations on Rank Histograms: Reliability of Temperature and Wind Speed Forecasts from Finescale Ensemble Reforecasts, Mon. Wea. Rev., 139, 295–310, doi:10.1175/2010MWR3129.1, 2010.

Persson, A.: User Guide to ECMWF forecast products, Tech. rep., ECMWF, 2011.

Sundberg, R., Moberg, A., and Hind, A.: Statistical framework for evaluation of climate model simulations by use of climate proxy data from the last millennium – Part 1: Theory, Climate of the Past, 8, 1339–1353, doi:10.5194/cp-8-1339-2012, http://www.clim-past.net/8/1339/2012/, 2012.

Toth, Z., Talagrand, O., Candille, G., and Zhu, Y.: Probability and ensemble forecasts, in: Forecast Verification: A Practitioner's Guide in Atmospheric Science, edited by Jolliffe, I. T. and Stephenson, D. B., pp. 137–163, John Wiley, Chichester, U. K., 2003.