Dear editor, dear referees,

We hope our response finds you well.

We thank you for your careful and encouraging evaluation of our manuscript. Before we provide a detailed response to your comments, we have to highlight one relevant change in the revised manuscript.

During reconsidering the single-best analogue reconstruction test in response to referee #2's question on the discrepancy between the single best analogue and the tolerance area analogue reconstructions, we noted a bug in the preparation of the simulation data.

The preprocessing of the model data into the input for the analogue search resulted in the introduction of invalid values (Not A Number) for one proxy location in one model grid which was used in three simulations. Specifically, this affected the data from simulations with MIROC-ESM (Sueyoshi et al., 2013: Set-up of the PMIP3 paleoclimate experiments conducted using an Earth system model, MIROC-ESM, Geosci. Model Dev., doi:10.5194/gmd-6-819-2013).

This resulted in the fact that simulation data from these three simulations could never be selected as valid analogues in our tolerance area approach but were included in the single best analogue reconstruction and chosen as analogues there because the distance algorithm discounted the locations without numerical information. That is, the single best reconstruction was effectively calculated using a reduced number of locations.

Because the simulations were never part of any of our main reconstruction tasks, we choose the following solution to this issue. We remove from the manuscript the simulations for which data was erroneously prepared, and we redo the single best reconstruction. We acknowledge that this post-hoc modification of the experiment description is not optimal.

We invite you, the referees and the editor, to comment on this issue. We also invite you to provide recommendations whether the post-hoc change in experimental design should be mentioned in the manuscript.

Please see the tracked changes or the manuscript file for the details of all changes. Below we also list the major changes to the manuscript.

On behalf of the authors

Yours sincerely

Oliver Bothe

List of changes

- Experiment design
  - As noted above in our letter to the editor and the referees we had to adapt the single-best experimental setup to account for a bug. Subsequently discussions of the results had to be adpated as well.
- Abstract
  - Minor change
- Introduction
  - Minor clarifications
- Analog method, assumptions, and data
  - Clarifications and modifications on text and figures in response to the referees' comments and suggestions.
- Results
  - Adapted description of single-best results to the changed setup.
  - Clarification of figure captions in response and beyond response to the referees' comments.
  - Minor modifications in response to the referees' comments.
- Discussion
  - Additions in response to the referees' comments.
- Appendix
  - Minor modification.

Editor

Dear authors,

Thank you for submitting a revised version of your manuscript, which has now been evaluated by the reviewers. Both reviewers agree on the significance of your contribution. However, they also suggest some additional changes to improve clarity. The second reviewer also remains doubtful about the methodology, but they think the approach you suggest is worthwhile exploring and they are hence favourable to sharing it with the community. Nevertheless, I do encourage to take their concerns seriously and address them in a revised manuscript. The first reviewer also asks for additional clarification and some rethinking of the conclusions.

Best regards, Lukas Jonkers

Response: Dear editor, thank you for your comments. Below we respond to the reviewers' comments and try to address their concerns. We particularly aim to consider the points highlighted in your comments.

Referee 1

General Comments

I think the manuscript has been substantially improved, especially with the clarification of the method and the inclusion of a better designed Pseudo Proxy Experiment. However, I still have some concerns regarding: (1) the clarity of the method's explanation (which should be central for a technical note), (2) some unsustained conclusions and (3) the repetition of several sentences along the text.

Response: Thank you for this positive evaluation. Below we try to address all your points in detail.

I indicate lines and pages as in the file with tracked changes.

(1): Method's explanation 1. In this version of the text the method is much better explained. However, I still can't find anywhere a description of what is meant by 90%, 99%, etc. uncertainty ellipses.

The first appeareance of the 90%, 99%, etc. in the text is in the following sentence (line 33, page 7):

"We can do so at different levels of proxy-time uncertainty, e.g., 90% or 99%, comparable to common expressions of uncertainty intervals."

At other time it says: "To do so, we consider the proxy values valid at all dates within their a 90% dating uncertainty, then identify the range of these values, and take the mid-point of the range as the proxy value for this date." Which, to my point of view makes no sense.

I think a clear definition of x% ellipse uncertainty need to be provided, as it is a main part of the methodology.

Response: We understand the referee's concern. We address this now. We provide a quick derivation of our percentage tolerance view from classical confidence intervals and the chosen language there.

We also recognize that we use percentages differently, in so far as we also use it for the coverage of the reconstructions. We do not think we can avoid this double usage.

However, we agree that we have to change the occurrence mentioned by the referee. First, the sentence is apparently grammatically incorrect. Second, we change the sentence to: "...we identify the range of possible proxy values within a certain dating uncertainty, and take the mid-point of that range as the proxy value for this date." We stress again here that this procedure is only used because we (i) consider proxies that are irregularly spaced in time but (ii) want to identify a single value for a date that (iii) allows us to compute a Euclidean distance.

2.

Line 23, page 17

"Valid analogues are those simulation fields that are within the resultant tolerance envelopes for all pseudoproxy locations available for a date."

I find this sentence of extreme importance for the methodology. However, it only appears in the Results section. I think it should be moved to the description of the method, where it is much needed. The term "valid analogues" appears several times in the text but I think in this sentence is the first time that a proper definition is given, thus its relevance.

Response: We account for the referee's comment by trying to add this information more clearly at more locations. We want to stress however, that it was explicitly mentioned in our method section.

(2): Unsustained conclusions 3. Line 11, page 29 "Generally, the reconstruction success appears to be better for proxy setups that only include UK0 records (Figure 11b,c)"

How can you say is „better“ when you don't know the truth? If you want to make a conclusion like this you should do this experiment in PPE setup. There is no base for such a statement.

Response: We appreciate the referee's concern and will clarify this. However, we feel forced to emphasize that our conclusion is not unsupported. By saying the success appears to be better we want to say that it appears that the reconstruction is more complete. We do not want to say anything about the quality of the resultant reconstruction in its relation to reality. We wrote so in the subsequent sentence. We change this now to: Generally, the method appears to provide more complete reconstructions among our proxy setups for those that only include $U_{37}^{K'}$ records (Figure 11b,c). That is, such consistent sets of proxies provide a more continuous reconstruction for both local tolerance assumptions.

Line 31, page 33

"That is, the method performs slightly better using 51-year averaged simulation data than using 101-year averaged data, and it performs even better using interannual data."

Again, I find this conclusions dangerous and erroneous, not derived from the present analysis. how do you know is better when you did experiments only when not knowing the truth (not PPE setup)?

I think this conclusion doesn't follow from your results and is inappropriate.

From you analysis you can only say that using 51 means leads to finding more possible analogues, if this. Whether this leads to a better reconstruction is not shown in this paper.

Response: The referee is correct in their last paragraph that this is what we can conclude. And this is exactly what we are stating in the paragraph.

We agree with the referee that we have to clarify this. However, again we want to stress that the whole paragraph from which the quote is selected deals with the reconstruction success in terms of finding analogues at all and not in terms of proximity to the truth. We indeed think the paragraph was written clearly already, but we modify it slightly.

(3): Repetition of sentences:

Several time in the text I found (almost identical) sentences repeated. Please, read the new text carefully and avoid this type of problems. Information should be more organized and not unnecessary repeated.

Response: We acknowledge the referee's concerns. We put effort into removing such duplications. We hope to have succeeded.

Here only 3 examples, but there are many more:

- Line 23, page 6:

We construct the pseudoproxies following the procedure of Bothe et al. (2019a, more specifically their ensemble approach)

Line 7, page 14:

We calculate the pseudoproxies following the procedure of Bothe et al. (2019a, more specifically their ensemble approach) but omit their effective dating uncertainty error term.

- Line 19, page 6:

The data is available in monthly resolution for the full simulation period for air temperature in 1.5 meter height, and as snaphots every ten simulation years for surface temperature

Line 18, page 14:

Data is available in simulation monthly resolution for air temperature in 1.5 meter height but only as snaphots every ten simulation years for surface temperature.

- Line 32, page 6

Figure 1a shows the 17 pseudoproxy locations and allows to identify their slight offset to the real proxy locations (Figure 1b) due to the discrete character of the simulation data

Line 1, page 16

Figure 1a shows the 17 pseudoproxy locations. These are close to the realistic proxy locations.

Response: We addressed the mentioned occurrences and hope to have addressed further problematic duplications.

Specific Comments

Figure 1: It is really hard to notice any differences between panels (a) and (b). Please, plot in only 1 panel and overlay a grid.

Response: We regard overlaying a grid as not helpful as it adds an unnecessary element. We now show the information from both panels in one panel.

Figure 3: Include 1 more column for P01 setup, in a different colour.

Response: We do so.

Figure 6 and 9: Panel (a) is in Kelvin while all the other panels and text are in °C. Please, unify.

Response: We do so.

Also, the captions are really unorganized and difficult to follow. The first describe all the panels and then go on to describe them all again.

Response: We tried to reduce redundancy.

Figure 7 and 10: Add latitude and Longitude.

Response: We do so.

You use „QUEST-Famous“ and „ QUEST Famous“. Please, select only one denomination.

Response: We thank the referee for noting this, we change the one occurrence of QUEST-FAMOUS to QUEST FAMOUS

Typo: snahots (appears twice in the text)

Response: We thank the referee for noting this.

Referee 2

General Comments:

Paper review

Title: Technical Note: Considerations on using uncertain proxies in the analogue method for spatiotemporal reconstructions of millennial-scale climate

Authors: Oliver Bothe and Eduardo Zorita

The paper describes an analogue method of deriving spatially-complete temperature fields from sparse proxy records by finding modeled climate fields which match the proxy data within uncertainty bounds. The paper first tests this method using a pseudoproxy framework and then does several experiments using real proxy data. The authors also consider alternate experimental designs, such as using different proxy records and different criteria for finding valid matches.

Overall, I think that this is worthwhile research, and it is always good to explore methodologies which can supplement our incomplete perspective on past climate from proxy records. Additionally, the paper is much improved from the first version. The authors do a good job discussing potential pitfalls or improvements to the method, which is useful since future improvements would benefit the methodology. The primary shortcoming of the paper is the fact that the method doesn't work as well as one would hope. Instead of finding insightful analogues for past climate, the method tends to either find too many analogues (resulting in a very broad estimate) or none at all. The authors acknowledge this shortcoming, which is good to see, and the title accurately reflects the paper's "work-in-progress" nature. As it is, I think this is a useful paper, with the understanding that the method presented does not appear to represent a final product but rather a stepping stone toward more successful methodologies in the future. Many of my concerns with the first draft, such as the somewhat confusing description of the methodology and unclear elements of the figures, have been improved. However, I still do not agree with the authors use of uncertainty ellipses. My comments and suggestions are discussed in more detail below.

Response: We thank the referee for their detailed assessment. Below, we respond to individual points.

General comments:

My general comments concern the overall design of the methodology. While the method described in the paper has clear limitations, I think that this is appropriately discussed to a large degree in the paper. The authors have tested, or at least discussed, some ways of improving the method in the future, which is commendable. However, I want to return to two of my comments from my first review:

1. First, the pass/fail dichotomy of the analogue search seems very strict. If one (or more) proxy has a considerable bias, this single record could cause the method to fail. This can be seen in the author's choice to exclude a proxy on p.11, stating that "we find (not shown) that including this record puts very strong constraints on the analogue candidates and can reduce the chance of finding valid analogues." The excluded proxy is "the alkenone unsaturation ratios of Bendle and Rosell-Melé (2007". This limitation could be mitigated if proxy uncertainty values were varied on a proxy-by-proxy basis (for example, by assigning a large uncertainty to certainty records individually, rather than increasing the uncertainty percentiles for all proxies simultaneously), but it may be difficult to find reasonable uncertainty values for each proxy. The paper decides to use consistent values across proxy records. To the paper's credit, the authors explore alternate experimental designs where analogues are found valid even if they fail to match a certain number/percentage of proxies. This is good. I'm not sure if anything else needs to be done regarding this point in the paper, but I would like the authors to continue considering this point if they pursue this research more in the future.

Response: We thank the referee for this comment.

We agree that future considerations have to include a thorough review of criteria when analogues are considered to be valid under uncertainty. That is, the method will benefit if we can find robust criteria that not only allow for uncertainty in the input proxy data but even explicit wrongness of this data.

As a side note: we are not sure that allowing for different tolerance thresholds among multiple proxies is a promising strategy as it may provoke criticisms of overly subjective assumptions. However, that does not imply that one should not try it.

In view of the referee's comment, for now, we do not add additional comments on this topic in the manuscript.

2. Secondly, I still feel that the shape of the uncertainty ellipses may lead to difficulties finding valid analogues. As I see it, the main problem with the use of ellipses is that the analogue search becomes stricter, rather than more lenient, for searches farther from the original date of a given datapoint. This presents a problem at the ends of records, and will only get worse as more proxies are included. For example, consider Figure 2b. As a thought experiment, imagine that the data point shown at 7.8 kYa BP is the oldest point in this proxy record. When searching for analogues at 7.8 kYa, any value from 22-30 degrees C would be deemed acceptable. However, when searching for analogues at 8.5 kYa (i.e. just inside the edge of the uncertainty ellipse), only a very small range of values around 25 degrees C would be considered acceptable. Why would the analogue search be stricter for a period where the data point is less likely to be valid? Shouldn't it be more lenient, since we don't know if the datapoint is even relevant at that age? Furthermore, if we look for analogues at 8.6 kYa (i.e. just outside of that uncertainty ellipse), any temperature would be deemed valid for this location (i.e. this data point provides no constraint on the analogues chosen for that date). This shift from somewhat strict (near the original age point) to extremely strict (at the border of the ellipse) to not-a-factor-at-all (outside of the ellipse) doesn't make sense to me. Wouldn't it make more sense for a data point to have less impact on the analogue selection for ages when it is less likely to be relevant? In the authors' response, they ask "Wouldn't we, in this alternative scenario, then overemphasize the ranges far away from the original dated age?" Yes, but that seems appropriate. The method would result in larger uncertainty bounds for times when the proxies provide less of a constraint. I'm not sure how this would be implemented mathematically in the methodology (perhaps through the use of rectangular uncertainty bands, as explored in the paper, to the authors' credit), but the current methodology is unsatisfying to me. One idea would be to have proxy uncertainties increase as one gets farther from central age of each proxy data point, and then require that the analogues be within the bounds of all proxy data points. The authors explain their reasoning for their choices in their response to my first review, but I disagree with their rationale. It is possible, however, that I am overlooking something.

Response: We acknowledge the referee's view.

Our argument for the ellipse is the following. We regard our time-date point as sampled from a two-dimensional distribution. If we regard this to be a uniform distribution, we would use a rectangular tolerance area. However, we regard the distribution as a two-dimensional Gaussian, which can be visualised as an ellipse in the two-dimensional plane. Thereby the probability density for a valid point reduces further away from the best estimate. If our analogue pool would well sample the climate space, we could weigh our time-data points by their likelihood within the two-dimensional Gaussian plane. Then values that are far off in either or both dimensions would be given less weight. However, as we have only a rather small candidate pool, we resort to a binary criterion of inclusion and exclusion.

I fear that the problems in both of the points above will become worse as more proxy records added. Users of the methodology could deal with this by excluding certain proxies (as mentioned in the example earlier), but this seems time intensive and potentially somewhat arbitrary. Regarding my second point above, it would probably take a lot of time to change the method entirely, but I would like to see a little more discussion (perhaps an extra paragraph) about this point in the paper.

Response: Once more we thank the referee for their thoughtful comments. We will add discussion on the second point.

Specific comments

- Page 1, line 19. By "climate state" do you mean the spatial climate field?

Response: We rephrase this now in this way.

- Figure 1 caption. Perhaps mention that the pseudoproxy locations are slightly offset from the actual proxy locations. I know that this is mentioned in the paper's text, but including a brief statement in the Fig. 1 caption should help prevent people from wondering why it appears that the same panel is shown twice.

Response: In response to referee #1, we now replot the Figure, and in response to referee #2 we also add a note.

- Page 6, lines 27-28. The statement "Using anomalies circumvents this issue" makes it sound like you are going to use anomalies, when you don't. Consider rephrasing to make this clearer.

Response: We thank the referee for pointing this out, our revisions try to make this clearer.

- Page 6, line 32: "Dansgaard-Oeschger (DO)" should be "Dansgaard-Oeschger (DO) events".

Response: We thank the referee.

- Page 7, line 5. I would change "original temperature units" to "absolute temperature units", since "original" makes it sound like you would be using the measured proxy units.

Response: We follow the advise.

- Page 8. Regarding the 90% and 99% uncertainty values (and later 99.9% and 99.99%), are temporal uncertainty values the same for each proxy?

Response: No. For the real proxies, we utilize the one data standard deviations provided in Marcott et al. (2013) for each proxy. Similarly, the pseudoproxy generation provides variable uncertainties. We clarify this.

- Page 9, line 17-18. The phrase "the envelope does not necessarily cover all years within the period of interest" doesn't make much sense to me. Can you rephrase?

Response: We try to do so, but also try to explain it here: the envelope stretches along the time dimension. The periods of interest is a continuous series of years from, e.g., 15,000 BP until today. The envelope, however, may be not continuous because of sample gaps.

- Page 9, lines 32-33: The paper states "... increase the probability to find a valid analogue at a certain date". But what if widening the time uncertainties made new proxies relevant to some ages? Couldn't that actual reduce the likelihood of finding an analogue (see page 24, lines 24- 26)?

Response: Exactly. We hoped to cover this case with "usually", but will be more explicit now.

- Page 12, lines 14-15. The paper says "Regarding proxy uncertainty, we try to identify as complete uncertainty estimates as possible from Marcott et al. (2013) and their references. For the sake of simplicity, we decided to assume an uncertainty of sigma = 1K for all proxies." If you use a common number for all proxies, why did you look at individual proxy values? Is 1K close to the mean or median value?

Response: The reason is that simply some references do not provide enough information to obtain a full uncertainty estimate. Our aim was to obtain full estimates for all proxies.

We modify the paragraph.

- Page 15, lines 1-2: I'm not sure exactly what this sentence means: "The latter modification also avoids that individual data points have an overly strong influence within our envelopes of tolerance." Can you rephrase?

Response: We try to clarify this.

- Page 15, line 4: I'm not sure what "black lines" refers to in this sentence.

Response: Sorry, we remove this now.

- Page 15, line 5: Remove the word "below".

Response: We do so.

- Figure 6, panel c: Perhaps rearrange the labels so that "QUEST FAMOUS target" and "pseudoproxy" are listed next to each other (since the pseudoproxy is generated from the QUEST model output, right?). Also, put the "Analogue median" and "Analogue range" labels next to each other since they're also related.

Response: We do so.

- Figure 6, panel d: The "Valid analogue examples" label only has one color, but the examples are shown in two different colors. Could you add the second color next to the label?

Response: We hope this is clearer now.

- Figure 6. Panels d and e took me a while to understand. If possible, could you try to explain it in a more intuitive way?

Response: We hope our modifications do now clearly describe the panels.

- Page 21, line 11: Is there any particular reason that age 8000 was chosen for the examples (or age 2430 for the examples in figure 9)?

Response: These years were chosen adhoc.

- Page 21, line 14. The paper says "Their local range at no point exceeds 4 degree Celsius". However, it looks like you are referring to the deviation from the median, rather than the total max-min range. Can you check this number and rephrase if necessary? Please do this for the "20 degrees Celsius" value a few lines later as well.

Response: The quoted numbers indeed already refer to the full local range.

- Figure 9, panels d and e: In these two panels, the two examples look like time-shifted versions of each other. Why is this? Is it because the same 101-year mean was chosen for each date?

Response: We thank the referee for this question. We now clarify in the text: "For the chosen year, there is only a small number of analogues, which form a sequence of consecutive simulated years from one simulation. Therefore, the two examples in panels (d) and (e) are simply time shifted sequences."

- Figure 11 caption: The caption says "All panels include the median, 90% interval, and full range for the reconstructions". However, only one range is shown. Is the displayed range the 90% interval or the full range? Please fix this line, both here and in the caption to Fig. 12.

Response: It is the full range. We apologize to the referees and the editor for this oversight. We correct both captions.

- Page 26, line 5: I would remove the comma after "Please".

Response: We do so.

- Page 26, line 6: Consider changing "further" to "also".

Response: We do so.

- Figure 12 caption. I would change "are valid if they fail" to "are valid even if they fail".

Response: We follow the referee's suggestion.

- Page 28, line 12: "warmer" than what?

Response: We clarify that we mean "warmer" compared to other setups.

- Page 28, lines 23-25: Consider discussing why the "single best analogue" differs from the default experimental setup (by large amounts for some ages). Is it because the "best" analogue may fail at one or more locations but still produce a better overall fit than other analogues? Since the "single best" methodology sounds similar to previous analogue methods, it would be interesting to explore why it diverges from your main experiment, and whether your method provides benefits over the "single best" analogue.

Response: We explore this now. While doing so we encountered a bug in our procedure that we also mention in our initial response to the editor and the referees. Therefore also the results for the single best reconstruction change.

- Page 30, line 2. I don't understand the use of "e.g." in this sentence. Please rephrase.

Response: We remove it.

- Page 30, line 11. What does "seasonal sensitivities" in this sentence refer to? I thought that you already considered seasonality in your proxy-to-model comparisons (as stated on page 12, lines 10-12). Or, if all proxy comparisons are made to annual-mean model data, than this needs to be corrected at several points in the paper (e.g. page 12 and Table 1 "season used").

Response: We clarify this. Seasonal sensitivities at this point do not mean the assumed seasonal sensitivities of our reconstruction process or a prior calibration but the true seasonal sensitivities of the recorders and how they may change over time. This includes that our assumptions may be too crude. We hope the following quote from Jonkers and Kučera (2017, https://doi.org/10.5194/cp-13-573-2017) for planktonic foraminifera clarifies our point: "seasonality changes with temperature in a way that minimises the environmental change that individual species experience".

- Page 32, line 30. The method succeeds only for some ages.

Response: We clarify this.

- Page 34, line 8. What does "central" mean in this sentence?

Response: Thank you. We mean by "central" here the part of the simulation name as listed in the respository that identifies which simulation was used. We clarify this by removing "central" and combining the two sentences.

Despite the quantity of these comments, I thought that the paper provides useful considerations and tests of the analogue method. Although I still disagree with the shape of the uncertainty ellipses, I think this is useful and interesting research.

Response: We thank the referee for this positive recommendation.