Review of “The Temporal Phasing of Rapid Dansgaard–Oeschger Warming Events Cannot Be Reliably Determined” by John Slattery et al., https://doi.org/10.5194/egusphere-2023-2496

The manuscript presents an analysis of some limitations of ramp fitting as a tool to determine the characteristics of abrupt events seen in palaeoclimate data and model runs. The study is relevant, generally well presented although (some more details could be wished for in places), and suitable for publication in Climate of the Past.

First, the ramp fitting method is extended to take pre- and post-transition trends in the data into account. This is a good move when the method is applied to model data which exhibit more pronounced trends in especially the interstadials, but see further comments below on the need for modelling a trend in the stadials. The main finding is that the ramp fitting method detects ramp start points that are generally too early when the method is tested on synthetic data, especially with high noise levels. This effect is denoted “bias”, and the bias is then quantified for a range of scenarios. The effect can be anything from one year upwards, depending mainly on the noise-to-signal levels.

I have some methodological concerns and some regarding the presentation of the results, that I think should be addressed before publication:

1. The manuscript does not really discuss the magnitude of the suggested bias compared to the credible intervals produced by the ramp fitting method as published in earlier studies. If, for example, the bias is a few years in cases where the credible intervals span decades, the bias will unlikely lead to wrong conclusions. In this case, the existence of a systematic bias would still be a valid result (and to people interested in the methodology also an important one), but the analysis should consistently address whether the bias is sufficiently contained within the existing uncertainty estimates.
2. There is often a downward slope during interstadials, especially in the model time series. However, it is not obvious whether the model should be changed to allow/detect a slope during the stadial pre-ramp period. Do the authors find significant (negative or positive) slopes of pre-ramp levels in real data (both proxy and model data)? It is simply not clear in the manuscript whether the addition of a slope in the stadial fixes a real problem. It is not surprising that allowing a negative slope in the stadial before a steep positive ramp will tend to lead to earlier ramp onsets, especially with noisy data. As I understand it, this would likely lead the authors to conclude that the model without stadial slopes has a bias. But if there is no significant slope in the stadial (and preferably a physical reason to why such a slope would be expected), it is not clear that allowing the model to have a sloping stadial is an improvement in terms of representing the underlying physics. If there is no significant stadial slope, an observed difference in the timing of the ramp onset between the two models (with/without stadial slope) is simply a result of different assumptions about the signal structure and not necessarily a bias as such. The authors should analyze this question carefully to justify whether/when/why addition of a pre-ramp slope is an improvement. And they should test (for synthetic data created without any pre-ramp slope in the noise-free data) how adding this extra degree of freedom changes the detected ramp characteristics and report the results explicitly in the manuscript.
3. In Capron, the results of a single test of possible bias are given in the Methods section (and Fig. S2). Here, no significant bias was detected (the 50-year duration of the synthetic data was reconstructed as approximately 51 years by the method (and as an add-on to the previous point: the results in the S2 figure illustrate that the uncertainties are considerable and easily would include a bias of a few years). Because the results differ, it would be interesting if the authors could discuss their analysis in the context of the published test. Maybe the way the test data were created could make a difference (also see next point)?

4. The authors do not describe in detail how they generate their synthetic test data, and when they estimate the bias, how many independent realizations are run (my experience is that especially for high noise levels, averaging over at least several tens of realizations is necessary). This information should be provided somewhere. Also, and especially because the conclusions differ from those of earlier tests (e.g. those mentioned in Capron), the synthetic data series used here should be made available to allow others to make future tests against the exact same test data, thus reducing the risk that differences could be caused by implementation issues.
5. Related to the former point, one could suspect that especially when adding relatively high levels of autocorrelated noise, the transitions (i.e., ramps) of the synthetic data actually become more gradual (with longer durations and/or earlier onsets as a consequence). This is what was tested in Capron Fig. S2 for one scenario, but the results could be different for other parameters and other noise-to-signal levels (see PS below). If this indeed is the case, I would say that it changes the way we should look at this: it's not "just" the ramp fitting procedure that has an issue, but an inherent limitation related to the nature of climate data with autocorrelated noise (or data affected by autocorrelated short-term variability, if you like). I think this issue is related to what the authors are hinting at around line 226-228, but it could be expanded and made clearer.
6. Somewhat along the same lines: It is not surprising in itself that high noise levels lead to longer transitions: As the noise level increases, (probably all) ramp detection methods will have a harder time reconstructing the (no-noise) ramp duration, and there will therefore be a larger spread in the detected ramp durations. If the noise is large enough, the range of ramp durations becomes asymmetric because the ramp durations by design are positive. An example with a synthetic ramp duration of 50 years: For high noise levels, the distribution of accepted solutions produced by the method will include values that are larger than 100 years, but by design none which are smaller than zero. The mean of the distribution will be biased towards larger durations and lead to earlier onset points (and/or later ramp end points). In the test described in Capron, this effect of asymmetric distributions is observed although it does not lead to a bias in transition duration similar to the one reported in the current manuscript (see PS below). However, by using the median rather than the mean, the effect on the result can be reduced. This highlights two points. Firstly, using the median may be more appropriate than the mean. This should be discussed and tested explicitly. Secondly, this is not really an issue with the ramp fitting method, but a consequence of noisy ramped data.

   The observation is consistent with Fig. 3h which shows that there is largest bias towards early onsets for the most noisy and fastest transitions.

I think it would strengthen the manuscript a lot if the authors could go a bit deeper into discussing the issues raised in point 5 and 6. If they agree with these points, I encourage them to think about how to phrase their results in a way that better acknowledges that it's not "just" a methodological problem with ramp fitting, but an inherent limitation that both is related to the method and to the nature of autocorrelated ramped data and/or that (some of) their bias comes from the way the synthetic data are created. The conclusion of Capron already discusses some of these issues (and tries to separate those that originate from the complexities of the climate system and those that are more related to the noise and method). It would be good if the authors spend a bit more space on discussing how their results confirm, refine (or even contradict) the ideas presented there, and separate between the different types of difficulties.

[PS on the test from Capron Fig. S2:
While writing the review, and with inspiration from the manuscript and some tests performed by a student in Copenhagen last year (not published), I first ran some more realizations of the same test as in Capron: The mean reconstructed ramp duration was 51.9 years, and the median 51.5 years.

This would probably correspond to an early ramp onset bias of ~1 year, considerably less than reported in the manuscript at hand, and well within any reasonable uncertainty estimate based on the width of the distributions. Reducing the noise level to half led to no significant bias (median 49.6 years, mean 50.2 years). Then I tried with double noise level (stadial noise level is the default in this test, which is already a relatively high noise level) and observed a more pronounced bias in transition length, which was also larger when using the mean rather than median. I did not check this through well enough to share the results, so I provide only this indication. I think this supports the idea mentioned in point 5-6. It also indicates that in Capron (at least for Calcium and in the absence of significant slopes of the pre- and post-transition values), bias does not play any important role for the results.]

On a more technical note, I assume that the authors have tested their implementation (including generation of synthetic data). It would be nice with a summary of the tests: Does the method faithfully identify the properties of the synthetic data (including the autocorrelation time) for low noise levels? Figure 3 goes some way in demonstrating this, but I think a bit more details and a dedicated section would be appropriate (possibly in a supplement).

Finally, I think the title is not fully aligned with the content and conclusions. The following line from the conclusion “Despite the bias that we have uncovered here, we nevertheless suggest that this Bayesian ramp fitting method remains the best approach to understanding the temporal phasing of DO events” does not really work well together with the title “The Temporal Phasing of Rapid Dansgaard–Oeschger Warming Events Cannot Be Reliably Determined”.

In summary, I think the manuscript is appropriate for publication in Climate of the Past, but that additions and revisions are needed to fully document the analysis and make it clearer to which degree and under which circumstances the extended model with stadial slopes represents an improvement in the context of the data at hand, discuss in more depth whether the bias is a methodological problem or arise from the inherent difficulties of identifying ramps in noisy data, and assess whether the proposed bias will influence earlier conclusions based on the ramp fitting methodology.

Review by Sune O. Rasmussen, Jan 10th, 2024.

Line-by-line comments

29: ”However, as of yet it has not been possible to conclusively identify a cause for these rapid Dansgaard–Oeschger warming events” could be rephrased. The exact mechanisms leading to the described changes are not understood but the AMOC state change is generally accepted as the main cause of most of the observed changes in the climate system.

34-39: The papers of Adolphi and Corrick align records from ice cores and speleothems and have to deal with non-climatic dating offsets between different dating approaches. The WAIS and Svensson papers use synchronized ice-core records. Because the uncertainties (in terms of both magnitude and nature) are quite different, it would be better to describe them independently. Bias in event detection of ~20 years are only marginally relevant in this context, which is not the case for the methods described in line 39-41.

39: (Buizert et al., a) should be cited as “WAIS Divide Project Members”.

47-48 and 51-52: Elaborate or remove. The current sentence is vague. How do they contradict each other (given the error bars)? Consider if it would make the discussion clearer to include a discussion of the difference of absolute vs. relative dating errors. An ice-core timescale can build up significant bias over long stretches, and the published maximum error of GICC05 is therefore large in MIS 3.

58-63: I believe the papers came out in the reverse order: Capron et al. was accepted while the manuscript of Riechers and Boers was in the discussion phase. Please adjust wording accordingly, and in particular, be careful to cite the conclusions of Capron et al. correctly. I do not find a statement in Capron et al. to support the claim "they also concluded that applying a method of the type developed by Erhardt et al. (2019) to ice core data may not allow the identification of a unique order". In contrast, Capron et al. acknowledges that the Erhardt et al. approach of stacking may be appropriate under certain assumptions. The lack of a well-established order of events which is similar between events could both be due to noise / internal variability making detection difficult, but also that such a "standard sequence of events" may not exist.

72: Maybe reconsider use of "cause" (as above).

90: Proofreading needed.

93-94: Vague. Refer to Vettoretti et al. 2022 Fig. 2a or describe how your analysis is different.

110: It is not clear what "parameter ranges" refer to.

117: Describe somewhere (here, in Methods, or supplementary material) what "for which the method was successful" means.

119-120: Add charge to ion or spell out full names.

120: Note that Capron et al. did not make a one-to-one "identification" of proxies and model data extracts: "Over each D-O-like transition, we extract the time series of four climatic measures from the model on the assumption that they reflect some of the same elements of the climate system as our ice-core proxy data". Revise accordingly.

122: Revise syntax.

137: It would be prettier if x(t) was defined for t_0 and t_1 as well (as in appendix).

154: This is a critical assumption which is also employed by Erhardt and discussed in Capron. I encourage the authors to consider and discuss the implications of the assumption (e.g. here or in section 2.5).

176: It is more common to use the SNR and not the "Noise / Signal ratio". I recommend using the more standard SNR unless there are good reasons for why not to do this.

178: Give more details on how the synthetic data were generated in order to realistically mimic true data.

192: (also see main point raised above). It is hardly surprising that adding a lot of noise to a steep transition tends to smoothen out the transition found by the ramp fitting method. In order to judge how serious this problem is, I suggest adding the following pieces of information:

- How does the calculated UMOTE compare to the distributions of onset times? I.e., are the confidence intervals wide enough to include the bias?

- It would be informative to add some estimates of noise/signal for the actual data used by Erhardt at the start of this section, so the reader gets a feeling for how the observed biases compare to what would be expected in real data.
- Same for the stadial and interstadial slopes: What are the realistic values for the Erhardt data sets?

2010-201: (Optional). Maybe comment on how an autocorrelation time on decadal scale could occur in the climate system?

202: Figure 3: Mention that the scales are not the same for a-d end e-h, respectively. I guess that noise/signal does not mean the same for annual and decadal data. Is there a way to compare the two more directly? For example for a certain annually resolved data set with a given noise/signal value, what is the corresponding noise/signal value for the same data set averaged to decadal resolution? That would elucidate the dependence of resolution more clearly.

206: For the tests in C1-C4: Are these tests run on the same synthetic data? How exactly were the synthetic data generated, and what value of the autocorrelation length was applied? Synthetic data with characteristics spanning a range of parameter values are tested, but without describing how the ranges compare to those of the real data, the results are hard to interpret in a useful way.
For the test on Figure C2: Is this relevant given that Capron did not use data with decadal resolution? Or am I misunderstanding the caption: “transitions with decadal resolution”?

214-216: Please provide the details of the autocorrelation range observed in the model data. What noise-to-signal characteristics were chosen? It’s fine to have the results illustrated in Fig. 3, but they should also be tabulated.

225: It would be clearer to integrate the section starting at 230 with this section.

241: Please explain what you mean by “reliability”.

245: Please provide table with the characteristics of the synthetic data that resemble the ice-core data sets.

265: (Figure 5). It is not clear whether the Calcium data have also been corrected for bias and how this factors into Fig. 5: Is the zero defined from corrected or uncorrected Calcium data ramps? Please provide the details.

303: I don’t see why this is “another issue”. It is quite clear that observing a too long duration is equivalent to the combination of a too early onset and a late end point, and as stated in the introduction, I would claim that both effects are likely to occur when you turn up the noise.

312-15: Section 3.2 is based on CCSM4-like data, and it is not clear if the same result holds for paleoclimate data. Please clarify/rephrase.

Throughout: Please add hyphens consistently to “sea-ice extent”, “ice-core records” and other similar compound adjectives. It’s somewhat a matter of style, but generally increases readability.