



AGU Advances

Peer Review History of

**Revealing the statistics of extreme events hidden in
short weather forecast data**

Justin Finkel¹, Edwin P. Gerber², Dorian S. Abbot³, Jonathan Weare²

¹Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology

²Courant Institute of Mathematical Sciences, New York University

³Department of Geophysical Sciences, University of Chicago

Files Uploaded Separately

Original Version of Manuscript (2022AV000749)

First Revision of Manuscript (2022AV000881)

Second Revision of Manuscript [Accepted] (2022AV000881R)

Author Response to Reviewers

Peer Review Comments on 2022AV000749

Reviewer #1

Review of "Revealing the statistics of extreme events hidden in short weather forecast data" by Finkel et al.

Recommendation: major revisions

This is an interesting approach. Better estimates of rare weather extremes are needed. The transition path theory looks promising. However, I feel it is poorly explained. Before I can recommend accepting this manuscript for publication it needs to substantially be improved.

1) I do not understand the way the clustering works. Why do you set the number of clusters to a high number (170)? Don't you get into estimation problems with such a high number of clusters and rather low number of data?

The whole approach should be explained in a more intuitive manner. The typical reader of AGU Advances is not a mathematician.

2) How do you deal with non-stationarities due to global warming? Your Markov model seems to be stationary.

3) In the introduction you mention that many weather extremes have a rather small spatial scale. However, SSWs are rather large scale. Hence, it is unclear how well your approach would work for smaller scale extremes which are localized while your SSW index is hemispheric. This should be discussed.

4) Line 259: You state that ERA20C is biased towards a lower number of SSWs, while in the overlap period with ERA5 they are comparable. How do you know this is a bias? Don't you implicitly assume stationarity here? There could have been systematic changes in the number of SSWs.

5) The manuscript would benefit from careful proof-reading. For example, lines 144, 338, etc.

Also some references need to be fixed, e.g. line 300 (Butler et al.).

Reviewer #2

(Comments begin on next page)

Review of “Revealing the statistics of extreme events hidden in short weather forecast data” by J. Finkel, E.P. Gerber, D.S. Abbot and J. Weare.

This review is done by Martin Jucker. I am revealing my identity because it would probably be obvious anyway, and in the hope that the authors can better understand my comments as they know my background.

Overall, the authors will find in my comments that I generally got confused, and I think the main messages are embedded within (possibly too) detailed explanations and comparisons. As a result, they are getting lost. I am not an expert in statistical methods, which might be why I didn't understand it, but if the manuscript is aimed at a broader audience within the climate community, I think my confusion is an indication that more work needs to be done on the way the results are presented.

Main comments

1. The Key Points only mention Transition Path Theory, not Markov State Models nor any of the other discussed methods. However, it seems to me that most of the manuscript is devoted to the MSM, while TPT is banned to the supplementary. As someone who does not know much about these methods, it was very confusing to read about so many other methods, sub-methods, similar methods, etc., such that I lost overview of what this manuscript really was about. For instance, in paragraph 170–193, the text uses the acronyms TPT, DGA, MSM, and LIM within just a few sentences. I fear I got lost.

What I would wish for is to move all of the methods which are not used in the actual analysis to the supplementary. That includes discussions of why those methods were not used, or how similar the current method is compared to others. As is, the important method, TPT, is discussed in the supplementary, and lots of irrelevant methods are discussed in the main text (apologies if I misunderstood this, but even so, it indicates a somewhat confusing structure of the paper).

Then, focus on just the method used here, and how MSM and TPT work together to form one coherent approach. The interested reader can then consult the supplementary for the details.

2. One of the main points the authors are trying to make is that they can assign occurrence probabilities to events which are too rare for the observational record by using a statistical method. However, what is done is to use the S2S dataset with all 10 perturbation members from the ECMWF model to construct a long dataset encompassing 900 years (even though, as the authors mention, the effective sample size is smaller). From this, the main question I have is this:

Is the better probability estimate due to using all members and therefore having more data, or is it thanks to a new statistical method? That is, could one simply use the basic method of counting events within the new, longer dataset and get similar results? I would welcome a discussion on this.

3. Related to above, I understand the authors are convinced their method can provide better probability estimates. But what isn't clear from the manuscript is: better than what? Maybe more importantly, where do the authors get their confidence about the estimates being better? It might be Figure 2, which I don't understand (see specific comments below), but again I think the important message here is buried somewhere underneath the details.

4. The S2S reforecasts try to predict the immediate future, and all of their simulations are therefore in some way linked to the real atmosphere. Thus, the 10 members of any given forecast are not independent in that they are all trying to predict the atmospheric state of that given year, including any particular phase of interannual and decadal variability. How can the authors be sure that the available period of 1996-2016 samples enough of the event space of the real atmosphere to be able to say anything robust about extremes?

Related to this, on lines 157-158, the authors state that “Many of them reach farther into the negative-U1060 tails than reanalysis, allowing us to calculate otherwise inaccessible probabilities.” But Figure 1a) shows that none of the S2S hindcasts go beyond the 2008-2009 U1060 from ERA5. Maybe consider showing an example of a strong SSW where the individual members produce an even stronger event.

5. The authors use different U1060 thresholds to detect more and more extreme events. They also extensively cite Horan and Reichler (2017) as those authors applied a different method, namely running thousands of years to get occurrence estimates, to try and fill out the sparse climate distribution. Why not apply the new method to the Southern Hemisphere where SSWs are truly rare (only one U1060 reversal on record)? This might of course be a somewhat personal way of looking at things, but it seems like a natural application, as the same long simulations used by Horan and Reichler (2017) were also used by Jucker et al (2021) to estimate the SSW frequency in the Southern Hemisphere. And of course, the one U1060 reversal happened in 2002, which is part of the data analysed in this manuscript. It would be very interesting whether this new method would be able to corroborate their results.

Jucker, M., Reichler, T., & Waugh, D. W. (2021). How frequent are Antarctic sudden stratospheric warmings in present and future climate? *Geophysical Research Letters*, 48, e2021GL093215. <https://doi.org/10.1029/2021GL093215>

Specific comments

Figure 2: I don't understand this plot. How can the points be outside their own error bars? The answer is probably somewhere in paragraph 120-229, but it didn't help me understand.

Paragraph 253-260: As written, this paragraph seems more a validation of ERA-20C than of TPT. Is this relevant to the message of the paper?

Probability current: This seems a bit self-fulfilling to me. Winters with SSW will necessarily show the arrows pointing towards the 0 line (∂B) as the current has to go through that boundary by definition. Again it's probably only the way it is discussed, but what is the advantage of using this compared to simply counting the number of SSWs for each day of the year?

L338: there's a Latex typo: “ $\sum \mathbf{Ject}$ ”

L377 and 427: As outlined above, I don't know how the authors can conclude the method is more precise. This is almost certainly linked to my misunderstanding of Figure 2 though.

Peer Review Comments on 2022AV000881

Reviewer #1

(Reviewer #1 did not provide comments)

Reviewer #2

I thank the authors for their efforts in responding to all of my previous comments. The manuscript is now much clearer and easier to understand. I particularly appreciate the added analysis of the committors and expected lead time, but I have two comments about these paragraphs (lines 454-503):

1. The text notes that a strong $U(t)$ is typically an indicator of an impending extreme SSW event. This is not that surprising and consistent with previous literature, such as Hocke et al. (2015) ([\[https://doi.org/10.5194/angeo-33-783-2015\]](https://doi.org/10.5194/angeo-33-783-2015)) (<https://doi.org/10.5194/angeo-33-783-2015>); Fig. 5) or Jucker (2016) ([\[https://doi.org/10.1175/JAS-D-15-0353.1\]](https://doi.org/10.1175/JAS-D-15-0353.1)) (<https://doi.org/10.1175/JAS-D-15-0353.1>) , see Figs. 5 and 8).

2. The section between lines 454-503 also highlights the potential of multiple variable combinations (lines 474ff). This is reminiscent of the work by Jucker & Reichler (2018) ([\[https://doi.org/10.1029/2018GL080691\]](https://doi.org/10.1029/2018GL080691)) (<https://doi.org/10.1029/2018GL080691>), which showed that the meridional PV gradient together with 100 hPa heat flux could predict the probability of SSWs in the future (although U and the PV gradient are related, they make physical arguments why the PV gradient might make more sense here due to its direct relation to the refractive index). Therefore, I wonder whether the dependence on U , $v'T'$, and their combination in this manuscript is similar to the one discussed in Jucker and Reichler (2018), and whether using PV gradient instead of U might increase the committor and lead time? I am not asking to re-do the analysis, but I think it would be good to discuss given that some choices have to be made among many possible variables and variable combinations.

Minor comments:

l100: I think it should either be "Figs. 1(a,b) show" or "Fig. 1(a,b) shows"

Eq (1): I would prefer using the actual number of weeks between Nov-Feb to get to the 900 years (which is 17.33333, so maybe it's not exactly 900 years) instead of 52 for the entire year and then dividing by 3. As I understood it, all other weeks of the year are never used, so showing the higher number of 2700 years might be misleading.

l171: "the" → "then"

l388: Again, I think this should be "Figs. 2b-e illustrate" or "Fig. 2b-e illustrates".

Figure 3, caption: The caption inverts the left and right panels.