



*AGU Advances*

Authors' Response to Peer Review Comments on

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

Justin Finkel<sup>1</sup>, Edwin P. Gerber<sup>2</sup>, Dorian S. Abbot<sup>3</sup>, Jonathan Weare<sup>2</sup>

<sup>1</sup>Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of  
Technology

<sup>2</sup>Courant Institute of Mathematical Sciences, New York University

<sup>3</sup>Department of Geophysical Sciences, University of Chicago

**Author Response to Peer Review Comments on 2022AV000749**

(Author response begins on next page)

# Reviewer comments and responses

Justin Finkel, Edwin P. Gerber, Dorian S. Abbot, Jonathan Weare

January 12, 2023

## General response

We thank Dr. Martin Jucker and the anonymous reviewer for their thoughtful and constructive feedback. The new submission has been substantially revised in four main ways.

1. In response to Reviewer #2's suggestion (2) to discuss the role of increased data in the reduced variance, as compared to specific details of the of the Markov state model (MSM) estimator, we have developed a simpler, more direct rate estimator compatible with the S2S data set. It that matches quite well with the MSM. We refer to the new rate estimation procedure as "flux-counting" and explain it in section 3.1.

While much simpler to understand and use, the flux-counting approach is restricted to rate estimates and, unlike the MSM approach, cannot be used to make predictions about an impending rare event. To emphasize this added value of the MSM method, we have now analyzed the results of the MSM calculations more thoroughly. As a result, two completely new figures have been added: Figs. 3 and 4 analyze the committor probability and expected lead time for SSW events as functions over state space. These are ingredients in the rate calculation which were only mentioned in passing in the first draft. We think the new results are more interesting and scientifically useful.

2. We have streamlined the methodological details of the MSM to a sequence of steps in section 3.2, which we believe is easier to follow than the presentation in the first draft. While we are using a Transition Path Theory framework, we feel this was overemphasized in the previous draft; reducing this part of the discussion makes it more clear how the MSM was constructed and used.
3. We have validated our algorithmic parameter choices more rigorously, not only by bootstrapping but by testing the MSM directly against data for estimating a "time-limited committor". This is explained in the supplement and we hope will be useful for any reasercher considering adopting the methods we present.
4. Given the addition of the new method and concerns from the reviewers, we've omitted the use of ERA-20C, to keep the figures and discussion succinct.

Specific revisions are detailed below.

## Reviewer #1

### General comments

**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.**

Response: We thank the reviewer for their interest in our work, but appreciate the concern that the method must be sufficiently explained before it can be of use to the geoscience community. As noted above, we have included a new, simpler method that we believe is more accessible. In additional, we have completely reworked the discussion of the Markov State Model approach, and sought to justify how this extra effort can provide more insight. We respond the specific concerns below.

## Specific comments

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.**

Response: The clusters are indeed motivated by the problem of limited data, but actually to help the situation, specifically, to lower the dimension of the problem. They allow us to reduce the phase space of the full IFS model (millions of variables) down to just 150 states. We chose a large number of clusters in order to represent a continuous function over state space, namely the committor probability, as accurately as possible.

We realize now that our method could have been confused with weather regime detection, which indeed calls for a much smaller number of clusters. The clusters are not meant to represent separate regimes, but just allow us to reduce the dimension of the space. Their exact structure is not important, but rather that they sufficiently cover the relevant dynamics. When computing the uncertainty bounds, we actually recompute them based on subsets of the data (leading to different clusters), to estimate the uncertainty associated with this dramatic truncation of the model space.

Lines 244-253 now emphasize this distinction and explain that the exact choices do not matter for the results:

...any number of clusters (denoted  $M_t$ ) from 50 to 150 gives similar results. In Fig. 2 we display results of a single representative choice of  $\delta = 5$  days and  $M_t = 150$ , along with a shaded 95% confidence interval derived from the pivotal bootstrap procedure [Wasserman, 2004] with 20 independent resamplings of the data (but without replacement). The supplement further explains how we selected these parameters to simultaneously optimize the MSM's fidelity and robustness on a simple performance benchmark. We emphasize that these clusters are not supposed to identify metastable weather regimes in the tradition of, e.g., [Michelangeli et al., 1995]; rather, they are a discretization of state space meant to represent continuous functions over that space, encoding gradual progress towards an SSW event.

As noted in the text above, we have now tuned hyperparameters more rigorously with bootstrapping and more extensive validation of the MSM's predictions on limited time horizons. This is explained in the supplementary material. As a result, we have slightly adjusted the cluster number and time delays from the first draft, but our results remain qualitatively — and quantitatively — the same.

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

Response: We do not study global warming explicitly in this paper, in part because the data come from a limited time span of 20 years, during which global warming's signature on SSW events is very small. (It is tenuous even with 60 years of observations, but does appear in centennial long climate integrations with strong anthropogenic forcing.) It is correct that our MSM approach would need to be modified if we were to consider a period long enough for global warming to have a significant impact on the dynamics. (The similar results from the flux counting method are in this sense encouraging, as this method would not be affected by trends.) We still believe our approach may be useful to assess global warming impacts, however, if the analysis were repeated on different 20-year periods. We point out this possibility in the discussion (line 542-551):

The rate we estimate from the S2S data set is based on 1996-2015 boundary conditions (sea surface temperatures, CO<sub>2</sub>), and our MSM method assumes the climate was stationary over this period. Our results indicate that according to the 2017 IFS, 1996-2015 conditions were more similar to 1959-2019 than direct counting of SSW events might suggest. This could mean that the IFS was missing some key climatological variable during that period [Dimdore-Miles et al., 2021]. There is, however, substantial uncertainty on the impact of global warming on SSWs, even under 4xCO<sub>2</sub> forcing [Ayarzagüena et al., 2020]. By repeating our analysis on different historical periods, or simulations initialized from climate model integrations under different forcing, one could discern a more decisive signal of forced changes than would be available from raw data.

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.**

Response: This is a good point, and we have added a note of caution in line 535-541:

Boreal SSWs provide an ideal demonstration of our method, providing both moderately and extremely rare events. A natural and intriguing future application is the rate of Southern-hemisphere SSW events, in the spirit of [Jucker et al., 2021], which is postponed to future work for the sake of brevity. The method may be extended to other kinds of extremes as well, though care must be exercised when defining the event (e.g., sets  $A$  and  $B$ ) and choosing features in which to do clustering (for the MSM approach), especially in the case of more spatially localized events.

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.**

Response: We have omitted the use of ERA20C.

We meant to say ERA20C had a slight negative bias during the same period. Therefore, yes, we implicitly extrapolated the bias to the rest of the dataset. Due to this and other issues, we have decided to omit ERA20C from the current draft, as ERA5 alone provides enough data to make the essential points.

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.).**

Response: We have proofread the manuscript more carefully.

## Reviewer #2

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.**

Response: We appreciate this feedback. To start, we have revised the key points to focus on the results. In an effort to draw attention to the essential parts of the method, we have restricted the jargon and distilled the calculations into a sequence of concrete steps. We now more clearly identify our (now second) method as an MSM approach, limiting discussion of the TPT framework. To acknowledge the mathematical background

and preceding literature, we follow with one explicit mathematical aside that the casual reader may skip (lines 332-361):

Let us take a brief aside to reference some mathematical context for the method above. The general framework that we have used to combine committor probabilities to compute rates and other steady-state statistics of rare transitions is *transition path theory* (TPT) [Vanden-Eijnden, 2014]. TPT has been applied to molecular dynamics [Noé et al., 2009, Meng et al., 2016, Strahan et al., 2021, Antoszewski et al., 2021], atmospheric and oceanic sciences [Finkel et al., 2020, Finkel et al., 2022, Miron et al., 2021, Miron et al., 2022] and social sciences [Helfmann et al., 2021]. Though TPT is typically formulated in a time-homogeneous setting, here we have built in explicit time-dependence to deal with the seasonal cycle, similarly to [Helfmann et al., 2020].

Our MSM-based approximation of the committor probability is similar in spirit to analogue forecasting [van den Dool, 1989], which is enjoying a renaissance with novel data-driven techniques, especially for characterizing extreme weather [Chattopadhyay et al., 2020, Lucente et al., 2022]. Dynamical Galerkin approximation (using a basis different than the one used here) and a short trajectory variant of analogue forecasting are tested on several benchmark problems in [Jacques-Dumas et al., 2022]. Formally, the transition operator encoded by the matrix in (6) is related to linear inverse models [Penland and Sardeshmukh, 1995], which have also been used to predict subseasonal extremes [Tseng et al., 2021]. Both MSMs and linear inverse models involve finite-dimensional approximations of the transition operator (or Koopman operator for deterministic dynamics) [Mezić, 2013, Mezić, 2005, Klus et al., 2018].

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.**

Response: While the simplest counting estimate is not correct for the S2S data set (because trajectories that do not cross the threshold within their 46 day lifetime may cross later), this question motivated us to develop a more direct rate estimate (relative to the MSM-based estimate) for SSW rates from the S2S data set, called “flux counting.” It is described in Subsection (3.1). The flux-counting estimate agrees very closely with the rate estimates from the MSM, which bolsters our confidence in the MSM. However, one can do much more with MSMs, e.g., one can study the committor probability and other forecast functions to learn about the statistical and dynamical structure of SSW progression. In the new version we have replaced our discussion of probability currents with an analysis of these functions. Subsection 4.2, “Statistical predictors of SSW”, is devoted to this. It begins as follows:

Estimates of long return times alone do not provide physical insight into the mechanisms driving the event. The committor probability and expected lead time estimates provided by the MSM approach encode information on the dynamics and predictability of SSW events, and on extreme events in general. These quantities cannot be computed by the flux-counting approach. A number of recent articles have pursued committor probabilities as windows into transitional dynamics, e.g., [Miloshevich et al., 2022] for European heat waves and [Frishman and Grafke, 2022] for the spread of turbulence in a pipe. On SSWs specifically, our own previous studies with a simple SSW model [Finkel et al., 2021, Finkel et al., 2022] found through sparse regression that a small set of physical variables could explain key variability in the committor.

We believe the new results and interpretation have strengthened the paper, both its motivations and its conclusions.

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.**

Response: The S2S estimates (both MSM and flux counting) are better because they have smaller error bars than ERA5. These estimates are thus much more precise, allowing one to establish whether a model (or different historic period) is statistically different. The tightening of the uncertainty bounds becomes increasingly important for the most extreme events, events that have only been observed a few times, or never at all. We've added text to the abstract, body, and conclusions of the manuscript to make this more clear.

The bootstrap procedure is standard, and summarized in lines 195-201:

Summing up these probabilities from Nov. 1 to Feb. 28, and sweeping over all thresholds  $U^{(th)}$ , we obtain the black curve in Fig. 2a. Error bars come from a bootstrapping procedure: we apply the estimate (5) to 20 different random 10-year subsets of  $\{1996, \dots, 2015\}$ , calculate the 2.5th and 97.5th percentiles of rate estimates, and form the pivotal 95% bootstrap confidence interval (see [Wasserman, 2004], chapter 8, for a formal account, although we have modified the procedure by sampling without replacement to maintain independence of different years.)

The whole MSM and flux-counting procedures are repeated for each 10-year subset. The smaller error bars from S2S data manifest most clearly in the very negative  $U^{(th)}$  range, where ERA5 has only a few events and therefore a huge relative uncertainty in the rate. Lines 205-209 state

...the error bars make clear that flux-counting enjoys a tremendous advantage over the direct ERA5 estimate. At all thresholds, the flux-counting error bar overlaps with the ERA5 error bar, but is much smaller. This gives us confidence to trust the flux-counting estimate farther into the tail where no ERA5 data are available.

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?**

Response: You're absolutely right that our conclusions are limited to the climate conditions of 1996-2015. As stated on line 542-544, "The rate we estimate from the S2S data set is based on 1996-2015 boundary conditions (sea surface temperatures, CO<sub>2</sub>), and our MSM method assumes the climate was stationary over this period." In other words, we are not trying to sample all phases of, e.g., the NAO and ENSO. Rather, we are augmenting the statistical sample that the earth gave us during that time period, to make a stronger statement about what *could* have occurred given the effectively random fluctuations in the atmosphere. Strictly speaking, we also take as truth the IFS's representation of that randomness, so in some sense our study is an analysis of the IFS itself. We devote a large part of the discussion section to these important concerns, acknowledging the need for extensions and comparisons (lines 522-534):

While the IFS model has proven outstanding in its medium-range forecast skill [Vitart, 2014, Kim et al., 2014, Vitart and Robertson, 2018], it was designed for short forecasts. It is not clear how it would behave if allowed to run for hundreds of years as a climate model, which requires careful attention to the boundary condition and conservation issues. Even if the climate were to remain stationary with its 1996-2015 parameters, numerical and model errors would inject some bias into the equilibrated simulation. Repeatedly initializing S2S forecasts with reanalysis ensures a realistic background climatology, and allows us to rely on the IFS strictly for the short-term integrations that it was designed for. Our method may be used as a diagnostic tool to compare different models against each other, with specific attention paid to their rare event rates. A useful extension of this work would be to repeat the analysis on multiple data streams from all 11 forecasting centers worldwide that contribute to the S2S project, as a different way to compare different models' ability to represent extremes.

5. **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.**

Response: Great suggestion. Fig. 1 has been upgraded to show one such simulated SSW event in December 1998, which is unprecedented in the observational record.

6. **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>.**

Response: This is also a very nice suggestion. Due to length constraints, we have decided not to pursue this question for the present paper. We began with the northern hemisphere in part because we wanted to validate our methods on less extreme events (SSWs that happen every other year) before pushing it to events that have rarely been observed. We have added this as a potential future application (line 536-537): “A natural and intriguing future application is the rate of Southern-hemisphere SSW events, in the spirit of [Jucker et al., 2021]”, and we are certainly interested in pursuing this idea.

## Specific comments

1. **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.**

Response: The original paper’s error bars actually use the S2S estimates as a “null hypothesis”, and were meant to answer the question “how likely were the real-world observations given the S2S probabilities?” In retrospect, this framing is unnecessarily confusing, and so we have now put error bars on ERA5 according to binomial confidence intervals, guaranteeing every point in ERA5 to fall within its error bar.

There is, however, one caveat worth mentioning in the case of the MSM error bars. We estimate those with bootstrapping, which means repeating *the entire clustering procedure* with various subsets of data. Clustering is a nonlinear, non-smooth function of data, which means there is no *a priori* guarantee that the rate estimate from all the data falls strictly between the lower and upper bounds of the rate estimate from subsets of the data. As a simple example, suppose our data consists of two scalars,  $x_1$  and  $x_2$ , and suppose we want error bars on the function  $f(x_1, x_2) = (x_1 - x_2)^2$ . Subsetting with replacement gives  $f(x_1, x_1) = f(x_2, x_2) = 0 < f(x_1, x_2)$ . In this case, bootstrapping does not give symmetric error bars but rather a systematic displacement. While this example is pathological, we see no reason that clustering could not also be pathological. In fact, for some parameter choices, the MSM bootstrap estimates are systematically displaced from the full-data estimate (see Fig. S2b). We believe this is a symptom of overfitting, and so ultimately select a parameter set that does not give this behavior (see Fig. S2a). This strategy is heuristic, but we think sensible.

2. **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?**

Response: We thought ERA20C needed some justification to be used as a longer-timescale comparison. But this dataset no longer plays a role in the current draft.

3. **Probability current: This seems a bit self-fulfilling to me. Winters with SSW will necessarily show the arrows pointing towards the 0 line ( $\partial B$ ) 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?**

Response: We have replaced the discussion of probability current by a more detailed analysis of the committor and expected lead time, which we think is more digestible and useful for this problem. Seasonal distributions are still displayed in Fig. 1, and show agreement between the MSM and flux-counting estimates.

**4. L338: there's a Latex typo: “sumathbfJect”**

Response: Thank you, this has been removed.

**5. 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.**

Response: The specific passage you reference is now removed. However, we have tried to clarify the advantages of S2S data in the response to your general comment (3) above.

## References

- [Antoszewski et al., 2021] Antoszewski, A., Lorpaiboon, C., Strahan, J., and Dinner, A. R. (2021). Kinetics of phenol escape from the insulin r6 hexamer. *The Journal of Physical Chemistry B*, 125(42):11637–11649. PMID: 34648712.
- [Ayarzagüena et al., 2020] Ayarzagüena, B., Charlton-Perez, A., Butler, A., Hitchcock, P., Simpson, I., Polvani, L., Butchart, N., Gerber, E., Gray, L., Hassler, B., Lin, P., Lott, F., Manzini, E., Mizuta, R., Orbe, C., Osprey, S., Saint-Martin, D., Sigmond, M., Taguchi, M., Volodin, E., and Watanabe, S. (2020). Uncertainty in the response of sudden stratospheric warmings and stratosphere-troposphere coupling to quadrupled co2 concentrations in cmip6 models. *Journal of Geophysical Research: Atmospheres*, 125(6):e2019JD032345. e2019JD032345 2019JD032345.
- [Chattopadhyay et al., 2020] Chattopadhyay, A., Nabizadeh, E., and Hassanzadeh, P. (2020). Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001958. e2019MS001958 10.1029/2019MS001958.
- [Dimdore-Miles et al., 2021] Dimdore-Miles, O., Gray, L., and Osprey, S. (2021). Origins of multi-decadal variability in sudden stratospheric warmings. *Weather and Climate Dynamics*, 2(1):205–231.
- [Finkel et al., 2020] Finkel, J., Abbot, D. S., and Weare, J. (2020). Path properties of atmospheric transitions: Illustration with a low-order sudden stratospheric warming model. *Journal of the Atmospheric Sciences*, 77(7):2327 – 2347.
- [Finkel et al., 2021] Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., and Weare, J. (2021). Learning forecasts of rare stratospheric transitions from short simulations. *Monthly Weather Review*, 149(11):3647 – 3669.
- [Finkel et al., 2022] Finkel, J., Webber, R. J., Gerber, E. P., Abbot, D. S., and Weare, J. (2022). Data-driven transition path analysis yields a statistical understanding of sudden stratospheric warming events in an idealized model. *Journal of the Atmospheric Sciences*.
- [Frishman and Grafke, 2022] Frishman, A. and Grafke, T. (2022). Dynamical landscape of transitional pipe flow. *Phys. Rev. E*, 105:045108.
- [Helfmann et al., 2021] Helfmann, L., Heitzig, J., Koltai, P., Kurths, J., and Schütte, C. (2021). Statistical analysis of tipping pathways in agent-based models. *The European Physical Journal Special Topics*, 230(16):3249–3271.
- [Helfmann et al., 2020] Helfmann, L., Ribera Borrell, E., Schütte, C., and Koltai, P. (2020). Extending transition path theory: Periodically driven and finite-time dynamics. *Journal of Nonlinear Science*, 30(6):3321–3366.
- [Jacques-Dumas et al., 2022] Jacques-Dumas, V., van Westen, R. M., Bouchet, F., and Dijkstra, H. A. (2022). Data-driven methods to estimate the committor function in conceptual ocean models. *EGUsphere*, 2022:1–35.

- [Jucker et al., 2021] Jucker, M., Reichler, T., and Waugh, D. W. (2021). How frequent are antarctic sudden stratospheric warmings in present and future climate? *Geophysical Research Letters*, 48(11):e2021GL093215. e2021GL093215 2021GL093215.
- [Kim et al., 2014] Kim, H.-M., Webster, P. J., Toma, V. E., and Kim, D. (2014). Predictability and prediction skill of the mjo in two operational forecasting systems. *Journal of Climate*, 27(14):5364 – 5378.
- [Klus et al., 2018] Klus, S., Nüske, F., Koltai, P., Wu, H., Kevrekidis, I., Schütte, C., and Noé, F. (2018). Data-driven model reduction and transfer operator approximation. *Journal of Nonlinear Science*, 28(3):985–1010.
- [Lucente et al., 2022] Lucente, D., Rolland, J., Herbert, C., and Bouchet, F. (2022). Coupling rare event algorithms with data-based learned committor functions using the analogue markov chain. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083201.
- [Meng et al., 2016] Meng, Y., Shukla, D., Pande, V. S., and Roux, B. (2016). Transition path theory analysis of c-src kinase activation. *Proceedings of the National Academy of Sciences*, 113(33):9193–9198.
- [Mezić, 2005] Mezić, I. (2005). Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325.
- [Mezić, 2013] Mezić, I. (2013). Analysis of fluid flows via spectral properties of the koopman operator. *Annual Review of Fluid Mechanics*, 45(1):357–378.
- [Michelangeli et al., 1995] Michelangeli, P.-A., Vautard, R., and Legras, B. (1995). Weather regimes: Recurrence and quasi stationarity. *Journal of Atmospheric Sciences*, 52(8):1237 – 1256.
- [Miloshevich et al., 2022] Miloshevich, G., Cozian, B., Abry, P., Borgnat, P., and Bouchet, F. (2022). Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *arXiv preprint arXiv:2208.00971*.
- [Miron et al., 2021] Miron, P., Beron-Vera, F. J., Helfmann, L., and Koltai, P. (2021). Transition paths of marine debris and the stability of the garbage patches. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(3):033101.
- [Miron et al., 2022] Miron, P., Beron-Vera, F. J., and Olascoaga, M. J. (2022). Transition paths of north atlantic deep water. *Journal of Atmospheric and Oceanic Technology*, 39(7):959 – 971.
- [Noé et al., 2009] Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L., and Weikl, T. R. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences*, 106(45):19011–19016.
- [Penland and Sardeshmukh, 1995] Penland, C. and Sardeshmukh, P. D. (1995). The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, 8(8):1999 – 2024.
- [Strahan et al., 2021] Strahan, J., Antoszewski, A., Lorpai boon, C., Vani, B. P., Weare, J., and Dinner, A. R. (2021). Long-time-scale predictions from short-trajectory data: A benchmark analysis of the trp-cage miniprotein. *Journal of Chemical Theory and Computation*, 17(5):2948–2963. PMID: 33908762.
- [Tseng et al., 2021] Tseng, K.-C., Johnson, N. C., Maloney, E. D., Barnes, E. A., and Kapnick, S. B. (2021). Mapping large-scale climate variability to hydrological extremes: An application of the linear inverse model to subseasonal prediction. *Journal of Climate*, 34(11):4207 – 4225.
- [van den Dool, 1989] van den Dool, H. M. (1989). A new look at weather forecasting through analogues. *Monthly Weather Review*, 117(10):2230 – 2247.
- [Vanden-Eijnden, 2014] Vanden-Eijnden, E. (2014). *Transition path theory*, pages 91–100. Advances in Experimental Medicine and Biology. Springer New York LLC.
- [Vitart, 2014] Vitart, F. (2014). Evolution of ecmwf sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1889–1899.

[Vitart and Robertson, 2018] Vitart, F. and Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *npj Climate and Atmospheric Science*, 1(1):3.

[Wasserman, 2004] Wasserman, L. (2004). *All of statistics*. Springer, New York.

**Author Response to Peer Review Comments on 2022AV000881**

(Author response begins on next page)

# Reviewer comments and responses

Justin Finkel, Edwin P. Gerber, Dorian S. Abbot, Jonathan Weare

February 25, 2023

## General response

Thank you to both reviewers for evaluating our revised manuscript. We are glad to hear our efforts made the paper clearer and more useful. We have added discussion in response to the minor comments of Reviewer #2, as detailed below. We also added a few words of clarification to the discussion of SSW rates (lines 363-363), to disambiguate between standard SSW events and extreme SSW events. The changes are in bold:

the S2S hindcasts recover the longer-term climatology, despite the (slightly) greater frequency of SSWs **of this intensity** from the period in which they were initialized... it does not appear that a differences in atmospheric boundary conditions (e.g., sea surface temperatures) caused a systematic increase in **intense** SSWs between 1996 and 2015;...

We have also made our code public in a Zenodo repository, as stated in the Open Research section.

## Reviewer #2

### General comments

**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):**

**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>); 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>) , see Figs. 5 and 8).**

Response: Thanks for this insightful comment. We agree that evidence of stronger winds before an event exist in the literature, both in the reanalysis (as far back as Charleton and Polvani (2007), which we cite as it predates the Hocke et al. study), and in modeling studies. We agree that Jucker 2016 is particularly relevant here, and have added Albers et al. 2014 as well. We have added a short paragraph to the discussion of the committor (line 465-471):

"Studies with reanalysis and idealized models [Charlton and Polvani, 2007, Jucker, 2016] have found a similar pattern of strengthening zonal wind, as well as meridional potential vorticity gradient, prior to strong SSW events. These effects are components of preconditioning, wherein the vortex develops a sharper edge and becomes more susceptible to the frequent upward bursts of wave activity emanating from the troposphere [Albers and Birner, 2014]. The presence of the same pattern in S2S is an encouraging signal of physical consistency across the model hierarchy."

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>)), 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.

Response: Thank you also for pointing out this paper, which has a similar spirit to ours. We have added a paragraph at the end of the section (lines 510-519) to point out the possibility to extend the regression to include more physically motivated features.

"Given the mature wave-mean flow interaction theory of SSWs, there are many other features likely to be as good or better at predicting SSWs. For example, from a long GCM integration, [Jucker and Reichler, 2018] found that meridional potential vorticity gradient and 100-hPa meridional heat flux—representing vortex preconditioning and wave activity respectively—can change SSW probability by roughly an order of magnitude at a one-week lead time, and still significantly at seasonal-scale lead times. At present, we have limited our regression analysis to features that are easy to compute without introducing noise by differentiation. But more specific physical hypotheses can be tested by enlarging the feature space to include the relevant terms. The same principle holds for other extreme events besides SSWs. "

## Specific comments

1. **I100: I think it should either be "Figs. 1(a,b) show" or "Fig. 1(a,b) shows"**

Response: We have changed it to "Fig. 1(a,b) shows".

2. **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.**

Response: Fair point. We have changed 900 to 875, which is the more precise estimate.

3. **I171: "the" → "then"**

Response: Thank you, we have corrected this.

4. **I388: Again, I think this should be "Figs. 2b-e illustrate" or "Fig. 2b-e illustrates".**

Response: It now reads "Fig. 2b-e illustrates".

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

Response: Thank you for catching this error; we have corrected the caption.

## References

- [Albers and Birner, 2014] Albers, J. R. and Birner, T. (2014). Vortex preconditioning due to planetary and gravity waves prior to sudden stratospheric warmings. *Journal of the Atmospheric Sciences*, 71(11):4028 – 4054.
- [Charlton and Polvani, 2007] Charlton, A. J. and Polvani, L. M. (2007). A new look at stratospheric sudden warmings. part i: Climatology and modeling benchmarks. *Journal of Climate*, 20(3):449 – 469.
- [Jucker, 2016] Jucker, M. (2016). Are sudden stratospheric warmings generic? insights from an idealized gcm. *Journal of the Atmospheric Sciences*, 73(12):5061 – 5080.
- [Jucker and Reichler, 2018] Jucker, M. and Reichler, T. (2018). Dynamical precursors for statistical prediction of stratospheric sudden warming events. *Geophysical Research Letters*, 45(23):13,124–13,132.