Author response to Reviewer 2

We thank reviewer 2 (Dr Leach) for his comments and thoughtful suggestions to improve the manuscript. In the following, we respond to each comment and describe the changes we have made to the manuscript based on them. Each review comment is reproduced below (in black), along with our response to the comment (in blue).

Line numbers refer to the 'tracked changes' version of the manuscript, which we will submit along with these responses.

Summary

The manuscript presents a forecast-based approach to attribution of surface extremes to sudden stratospheric warming events, suggesting and estimating a variety of metrics for measuring the strength of the link. This link is based on a comparison of free-running and nudged experiments in a variety of forecast models. The breadth of the experiments performed and resulting analysis is impressive, and provides a reliable counterfactual simulation approach and statistical framework for assessing the role of the stratosphere in individual weather events. My overall suggestions for improvement largely relate to the discussion of the GEV, the utility of the metrics produced, and the overall interpretation of the results. I have split my review into three sections: "significant suggestions", "minor remarks", and "other questions and ideas". The "significant suggestions" are broad areas that I believe the manuscript would benefit from addressing. The "minor remarks" are small suggestions relating to Figures or wording that will not substantially change the text. The "other questions and ideas" are comments that, in my view, are not necessary for the publication of this manuscript, but I hope may be useful and/or interesting for the authors regardless. Overall, I believe that this manuscript is suitable for publication with a few changes and additional discussion. For visibility, I have selected "major revisions" as in my view some of the analysis should be repeated under slightly altered definitions of the metrics, but would view the changes necessary to be at the very bottom end of "major", as hopefully is clear from my review.

We thank Dr Leach for his interest in our results and methodology, and for the valuable comments below.

Significant Suggestions

Statistical inference - on the use of GEVDs and derived metrics

Your acknowledgement of the (in)appropriateness of the GEV is appreciated, and I like your description of it as a regularisation tool, rather than the "truth" based on the estimated underlying distribution. However, I think it may be worth emphasising that the confidence intervals estimated are very likely underestimates (especially in the tails, which are at times pretty poorly captured). I think it's also important to point out that - besides the seasonal cycle - the assumption of identically distributed variables is only justified if the underlying processes are sufficiently similar. This is unlikely to be the case here, and in my view may

well contribute to the structure not captured by the GEV in figure 5 (ie. if the underlying distribution is a mixture). This lack of physical process understanding is particularly relevant for the snow depth analysis, where I would argue the GEV estimates are so poor that they should not be included. Particularly for CNRM, it seems likely that in the "free" ensemble the snowfall in those two top ensemble members was driven by sufficiently different processes that they are effectively drawn from a different distribution. While not necessary for this work, in my view a focus on the process understanding (which is a key advantage of forecast-based approaches over statistical ones) would be worth mentioning as an avenue for further research.

We agree that GEV confidence intervals will likely be underestimates and have added this caveat to the text (L204-205). We have also added a more explicit statement to the description of Fig 5 that the underlying distribution of the ensemble is likely to be a mixture representing different physical processes, so will not be fully captured by a single GEV. Regarding the GEV fits in the snow depth analysis, these are only shown to illustrate the poor fit and therefore to motivate the non-parametric approach taken in the snow attribution. We have made this more explicit in the text (L399).

We also agree that a physical process analysis would help the interpretation of results, including model differences. We reference several companion SNAPSI papers, some of which are now published (e.g. Lee et al. 2025) and some in prep/submitted (e.g. Kim et al. 2026), that give detailed process analyses of stratosphere-troposphere coupling within these experiments. While, as the reviewer states, a physical process analysis would be beyond scope for our paper, we hope that our results can be interpreted alongside these other publications.

The metrics used are pretty standard in attribution, though their specific definition as used here has at least one quirk, which I think is worth either addressing or noting. The risk ratio (RR) is based on the "observed" value, while the quantile shift (QS) is based on the equivalent value at the same quantile as that estimated from the ERA5 distribution. When the free distribution differs significantly from the, this can lead to the two metrics being calculated on very different "parts" of the distribution, leading to unintuitive results (such as the estimated attributable influence being an increase in probability but a reduction in intensity). I think it would be best to stick to using one "reference" to present the results: either calculate the quantile shift using the quantile that the observed value lies at in the *free* experiment (not in ERA5), OR calculate the risk ratio based on the value in the *free* experiment that corresponds to the ERA5 quantile. I believe that this would be consistent with how these metrics are typically used in the wider extreme weather attribution literature (considering the "free" experiments to be the "factual" experiment in attribution literature).

Motivated by this comment, we have modified the definition of relative risk as follows: we (i) measure the exceedance probability of the year in question relative to all ERA5 years, (ii) find the corresponding quantile in the free, early-initialized ensemble, and (3) measure all relative risks at this temperature, rather than the observed ERA5 temperature. Heuristically, this considers a given model's (*free, early*) distribution as a "model climatology", against which all forced and later-timed experiments should be measured. Ignoring the overall error of the (*free, early*) ensemble with respect to ERA5 is a form of bias-correction that focuses the results on comparing the different forcing protocols to each other, treating the dynamical

model as a “fixed effect.” The results are qualitatively unchanged, especially the case study of IFS, whose (*free, early*) ensemble matches climatology remarkably well. However, the multi-model ensemble results of Fig. 6 become stabler and more consistent, especially SSWJan2019 and SSWSep2019. We don’t do a separate quantile mapping for early and late ensembles of the same model, as the initialization date is an effect of interest in its own right. This comes at the expense of persisting degeneracy in relative risk (i.e. RR close to unity) in SSWSep2019. We feel the tradeoff is appropriate. The new approach is described over lines L250-271.

We suggest that RR and QS focusing on different parts of the distribution is a feature, not a bug, of our analysis. They are an attempt to distill a complex curve into a couple of numbers, and the occasional scenario where RR and QS vary in the counter-intuitive way — for example, CNRM-CM6-1 in the late initialization, where the nudged and control ensembles change shape drastically from early to late initializations, and cross at a particular place (see Fig. R1, below) — is worth acknowledging.

Additionally, an issue with the suggestion to use the free quantile of the observed value for QS, or the free severity of the observed quantile for RR, is that these values do not necessarily exist (or may be far in the tails) in the free GEV distribution, particularly in the presence of model bias (an example of this can also be seen in Fig R1). We think that our new method goes some way to making our metrics more internally-consistent, while also retaining some robustness to model bias.

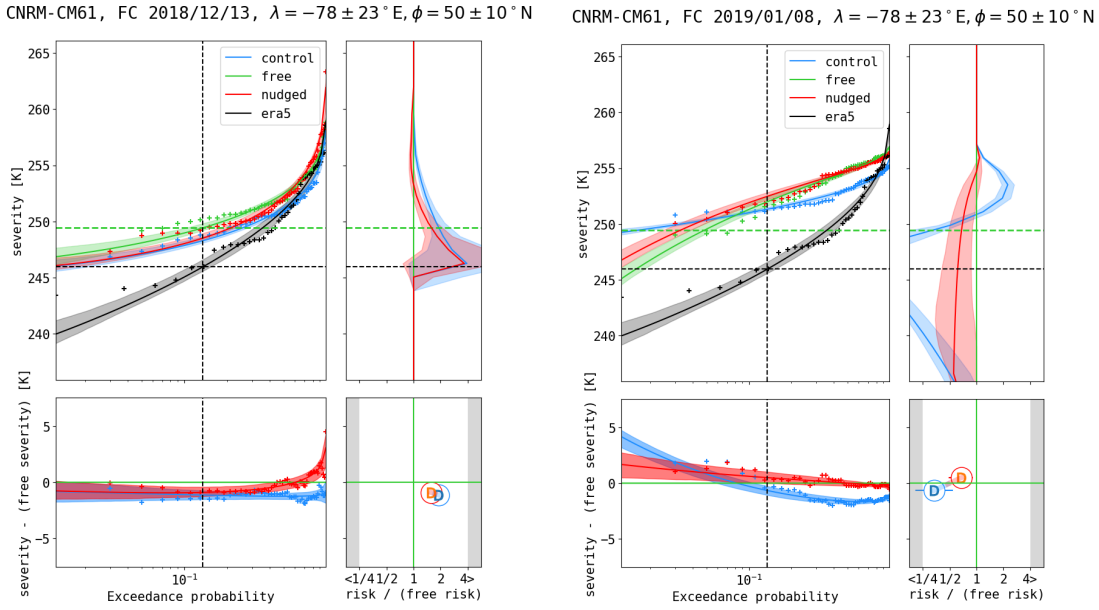

Fig R1. As Fig. 5 of the main text, but for CNRM-CM6-1 and SSWJan19.

My final point is along similar lines to the process understanding that I have mentioned. I think it's worth noting that the statistical analysis can only get you so far, and may actually misrepresent the true response, especially at the tails. This is particularly true for the CNRM snow depth one, where it is plausible (though I think an increased ensemble size would be needed to determine as this inference is based on only a couple of members) from the

plotted ensembles, that the nudging actually reduces the probability of truly extreme snowfall.

As we note above, we have managed some linking to processes by comparison with the weather regime analysis of Lee et al. 2025. However, we acknowledge that a more detailed connection of physical processes to simulated extremes would increase the robustness of our statistical analysis. We have added text stating this (L503-504).

Model bias and impacts on attributable influence

While you do include a section on the importance of model bias in the discussion, I think it may be worth making more of this in the results. In particular, model drifts could explain some of the discrepancies between the influence estimated in the long and short lead times and are not discussed.

In the case of the SSWSep19, I think that there are several potential issues that could be made more of or described in more detail:

- The seasonal cycle over the long lead date. From FigS5, it is clear that the majority of the ERA5 events are taken from the end of the window - which isn't available in the long forecast. This is a bit of a challenge as it introduces a bias that the RR metric will be sensitive to. This is a place where it may be more reasonable to define the event through a quantile for BOTH the RR and QS metrics.

We acknowledge that the truncation of the early initialization does lead to a problem for SSWSep19 (the only event for which both initializations do not cover the whole period of interest). We discuss this issue in L454-455. While, as we describe above, we have not used the same quantile for both RR and QS, our new approach does improve this issue, as can be seen in the improved consistency in Fig 6c.

- Similarly, in the short lead time, the RRs are all very small because the extreme event appears to have been entirely predictable (possibly because the models, IFS at least, have a slight warm bias too). This means that the event lies right at the bottom of the distribution of all the forecast experiments and thus the RR will be very small, even if the QS were large. Similarly to the long lead, this would be somewhat alleviated by consistently defining the event through a quantile rather than an observed value. I note that this may not make much difference.

We also agree with this interpretation, and acknowledge (as described above) that the non-overlapping periods for SSWSep19 limits our analysis for this event. Specifically, the fact that the (early, free) ensemble extremes are relatively moderate compared to those in ERA5 (due to its early truncation in the warming seasonal cycle) means that almost all late initialization ensemble members exceed the identified threshold, leading to RR near to unity (the warm model bias also contributes to this). While we could use a different methodology for this event (for instance, using (*late, free*) as the ‘climatology’ for late initializations), we are concerned this would add too much methodological complexity. We have adapted the text (L367-368 and L454-455) to better reflect this reasoning, and focus largely on QS for

SSWSep19, since this does not suffer from the same problem (highlighting a benefit of using both metrics).

Interpretation

It would be useful for you to provide some additional discussion on the interpretation of the metrics. In particular, it would be very useful to set out how the estimated influence could be dependent on the predictability of the event in the forecast models, and thus how readers should interpret the metrics given (especially the RR metric). This would give potentially very useful context for the snow depth analysis, where the estimated influence could be extremely dependent on the skill of the individual models at predicting extreme snowfall. The noted lack of model hindcasts for bias correction or calibration is a significant drawback here, as I think properly assessing model bias would strengthen the analysis.

We have added text after the introduction of RR and QS (L277-281) to explain the idea that RR should be considered with the underlying forecast skill in mind. Specifically, we state that, in cases where the free forecast is very skilful then risk is unlikely to be increased by nudging. We also discuss the role of predictability in the results, e.g. for the two initialisations of SSWFeb18, in Section 4. We acknowledge that bias correction would strengthen the results and dedicate a paragraph in Section 4 to this (L486-491). Unfortunately the only robust way to do this is with corresponding historical hindcasts with the same model versions as SNAPSI, and these were prohibitively expensive to request of modelling centers. Nonetheless, we suggest our new RR calculation method does go a little way toward bias correction, as described above.

Minor Remarks:

- Fig 3: it would be easier to see differences if the IFS panels were shown as differences? At the moment they look much the same and so the IFS panels don't really provide additional information.

We have changed the right hand column of Fig 3 to show differences between IFS and the ERA5 climatology as suggested.

- L186 doesn't make sense? Is the min / max over the 4 daily samples relevant given that the min / max is computed again over the whole integration? Or is this just for clarity to turn the 6-hourly samples into daily estimated minima or maxima?

This is the correct interpretation. The intensity, S, is the daily maximum from the 6-hourly samples, while the severity, S*, is the maximum intensity over the relevant time window. This has been clarified in the text (L188-189; also in response to a similar comment by Reviewer 1).

- L201 on the independence -> may also be worth including that the presence of skill means that independence of the maxima of the blocks is violated (in addition to the independence of the variables within the blocks).

We have changed the example of the violation of independence to “unlike an ensemble of skillful forecasts” as we agree this is a more relevant example than consecutive days (L204).

- L213 -> there is a growing body of work into physical bounds on the limits of temperature extremes (and some work into using these physical limits to constrain statistical models of such extremes). May be worth mentioning. I have provided some references below.

We have added these and one more reference – thanks for pointing this out (L223-225).

- Fig 4 / 5: the x-axis scales used cut off data points in 4 and make them nearly invisible in 5.

The x-axis scale has been changed to fix this issue.

- Fig 4: in (i), are the "GEV" fits the median of the bootstrap sample of fits (as opposed to the GEV fit to the original data)? Just asking as the solid black line has a kink around an exceedence probability of 0.6-0.7 which looks a little odd in the context of a parametric GEV model. Or is this because a linear scale without sufficiently small spacing is used to produce the line?

This was an artifact of the linear scale. We have refined it and the curves are smoother now.

- Fig 4: Could a different symbol be used for the "observed" sample? The bigger cross is hard to pick out. To make it even more clear, the horizontal reference line could be drawn across both panels (i) and (ii)

The symbol for the observed sample in Fig 4 has been made larger, and changed to a circle.

- L255: "dots" -> "crosses"

This has been fixed.

- L240: I'm a little unsure about using the "absolute risk" terminology, given that the probabilities estimated throughout are conditional in at least some sense.

We have chosen to keep the term “absolute risk” in order to clearly differentiate it from “relative risk” and also that a consistent term can be used to apply to both ERA5 and models. We acknowledge that forecast risks are conditional on initialization, however, and we have added the following to make this clear “noting that, although we use the term `absolute', AR remains conditional on the forecast initialization and any nudging” (L258).

- [LEFT IN FOR TRANSPARENCY - though I think this is to do with the quantile-based vs. observed baseline issue I've commented on above] Fig 6: One feature I am struggling to understand is how figure 6 appears to show a reduction in risk for an increase in the severity of the extreme assessed in some cases (eg. panel c/I nudged B)? Could you comment on how this is possible? I wondered if this was something to do with how the confidence intervals are plotted, but given that this feels non-intuitive, this it's worth commenting on.

The example pointed out has been made more consistent by our new RR methodology, meaning that models in Fig 6c fall along the 'expected' relationship between RR and QS (being along the upper left-lower right axis for cold spells and lower left-upper right axis for heat waves). However, one model in panel b(ii) (D - CNRM-CM6-1) does fall outside this relationship and is the example shown in Fig R1. In this case the unexpected outcome can be understood in terms of a change in the shape of the GEV distributions. We comment on this possibility, arising from the nonlinear nature of GEVs, on L475-479.

- Fig 5: explain the dashed lines (these are the ERA5 observed value and quantile per the text, but it would help to state this in the caption or somewhere on the figure).

An annotation added to the figure 5 in order to explain this (as well as a new green line, showing the free equivalent risk quantile used in our new approach to RR).

- L265-270: is it worth stating that the equally the free experiment could be expected to better match climatology for a low-predictability situation given that the free experiment has the additional dispersion in the stratosphere and is only conditional on the predictable component of the weather at the point of initialisation.

We agree with this point. We include in the text the following, which we hope makes clear that it either free or control might be expected to match climatology better: "however, nudging toward a single mean value does not imply matching the full distribution better, and might equally be expected to reduce the variance across members as a result of the lack of dispersion in the stratosphere and its influence on the troposphere. In other words, it is not obvious whether free or control experiments should better match the climatological distribution" (L296-299).

- Fig S6 (a)(i) and (a)(iii) appear to disagree - all the red crosses lie below the green in the tail (eg. less likely that 10^-1) of (i), but in (iii) the severity change indicated by the crosses is that the nudged severity is higher? I wondered if this is a sign error, but I don't think so since the lower part of the distribution looks consistent.

Thank you for catching this error. Indeed, we were plotting the crosses in the wrong order, not properly accounting for the opposite sign of extreme temperatures in the austral case. It has been fixed now.

- Is it possible to include a description of the model specifications (model cycle, resolution, ocean etc.) in the supplement? I think this would be useful.

We have modified Table 2 in the main text to include this information (where it is possible to fit, without creating another table), in addition to relevant references for each model version.

- An interesting feature relevant to the discussion between L263-275 is that the ensemble variance relative to the free increases in not only the control but the nudged experiment too (per Fig 4; though the differences in variance may not be significant between the free and nudged experiments). This feels unintuitive - can you explain why?

We agree that the larger GEV scale parameters (sigma values in Fig 4) for nudged and control relative to free might be unexpected, given that the evolution of the system is more constrained in these experiments. However, we would also stress that, in a coupled nonlinear dynamical system, it is not necessarily the case that constraining the variability of one part constrains that of another.

We have added standard errors to the GEV parameters in Fig. 4. The only significant difference (in terms of errors not overlapping) is that between free and control for the later initialization. We suggest this is because the control has a greater number of ensemble members with relatively warm cold extremes (as can be seen from its risk function extending to higher temperatures in Fig 5). This is also consistent with a larger number of NAO+ members in control relative to free, found by Lee et al. 2025.

Other Questions and Ideas:

- Would a "better" (or perhaps actually just "different") experiment design for the control be to nudge each ensemble member towards a different (random) realisation from the observed climatology? This would provide internally physically plausible/realisable stratospheric realisations with the same stratospheric variance properties as the observed realisations, rather than the nudging towards the climatological mean?

Yes we completely agree with this suggestion and have added it to the discussion (also in response to a similar comment by Reviewer 1) (L497-498).

Regards,

Nick Leach

References

Noyelle, Robin, Yoann Robin, Philippe Naveau, Pascal Yiou, and Davide Faranda. 'Integration of Physical Bound Constraints to Alleviate Shortcomings of Statistical Models for Extreme Temperatures'. Journal of Climate. Journal of Climate 1, no. aop (2026). https://doi.org/10.1175/JCLI-D-25-0112.1.

Noyelle, Robin, Yi Zhang, Pascal Yiou, and Davide Faranda. 'Maximal Reachable Temperatures for Western Europe in Current Climate'. Environmental Research Letters 18, no. 9 (2023): 094061. https://doi.org/10.1088/1748-9326/acf679.

Zhang, Yi, and William R. Boos. 'An Upper Bound for Extreme Temperatures over Midlatitude Land'. Proceedings of the National Academy of Sciences 120, no. 12 (2023): e2215278120. https://doi.org/10.1073/pnas.2215278120.

References in this response:

Lee, R. W., Charlton-Perez, A. J., and Lee, S. H.: Stratospheric impacts on weather regimes following the 2018 and 2019 sudden stratospheric warmings, Geophysical Research Letters, 52, e2025GL115 668, https://doi.org/10.1029/2025GL115668, 2025

Kim, H., Butler, A., Garfinkel, C., Hitchcock, P., Hong, D.-C., Rao, J., Lawrence, Z., Anstey, J., Ayarzagüena, B., Baldwin, M., Charlton-Perez, A., Erner, I., Henderson, S., Hu, D., Hyun, Y.-K., Jia, L., Kang, M.-J., Karpechko, A., Kharin, V., Knight, J., Koren, G., Lang, A., Lim, E.-P., Lin, H., Lee, R., Lee, S., Malguzzi, P., Manney, G., Mastrangelo, D., Muncaster, R., Paquette, C., Park, C.-H., Polichtchouk,I., Polvani, L., Richter, J. H., Seviour, W., Sigmond, M., Simpson, I., Son, S.-W., Specq, D., Stockdale, T., Taguchi, M., and Xiang, B.: Quantification of stratospheric influence on surface predictability: An overview of SNAPSI results, in prep, 2026.