Author response to Reviewer 1

We thank reviewer 1 for their 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.

In this work, the authors use the SNAPSI experiments to attribute the role of the stratosphere in the weather extremes that followed 3 sudden stratospheric warming: two in the NH and one in the SH. The forecast-based attribution approach is well motivated, and the use of the relative risk (RR) and quantile shift as complementary metrics is a good approach to complement their weaknesses. The paper is clear and well structured and it would be a good contribution to the journal and the studies done with the SNAPSI experiments. The main areas that could be strengthened are the definition of severity and the attribution when RR becomes ill-conditioned, potentially by adding the Extreme Forecast Index (EFI) and/or shift of tails (SoT) diagnostics. See my comments below

We thank the reviewer for their interest in our analysis.

Major comments:

Definition of severity: The paper discusses minimum temperature as the relevant extreme for 2 SSW cases, but the severity definition equation is written as a maximum over time. This is consistent if the S is defined with a sign convention (eg. coldness = -T) or if the variable is transformed prior to the maximization (which I don't think is the case if I read correctly the previous two paragraphs). At present this can be a bit confusing. Could you clarify this in the eq. 1

The intensity, S, for SSWFeb18 and SSWJan19 is calculated, as the reviewer suggests, by first multiplying the temperature for these events by -1 and then finding maxima (such that a more positive S represents a colder temperature minimum). While for SSWSep19 we do not multiply by -1, and a more positive S represents a warmer temperature maximum. In this way, the maximisation to define severity, S*, in equation 1 applies to all three events. We have provided this more explicit explanation of the methodology in the text (L188-189).

Using RR in unstable cases: RR based on exceeding the observed severity can become 0 or infinite or extremely unstable when the event is too rare for the model. You already discuss this in the text and partly address it by reporting the quantile shift. The EFI and SoT would be valuable additions because they remain informative when the observed threshold is outside the ensemble outcomes and it is used in operational context for the s2s timescale. Could you add this type of diagnostic to your study? Could be interesting to see the changes in EFI/SoT for each experiment and possibly report the differences. EFI and SoT are used, for instance, by the ECMWF to predict extreme/unusual events and and the results from this

analysis could also be of interest for prediction centers if they are presented using a quantity already in use by them. You may end up with the same weaknesses as in the study in terms if models have strong bias but it can still be useful.

We thank the reviewer for this suggestion. We have subsequently investigated whether the Extreme Forecast Index (EFI) and Shift of Tails (SoT) metrics could be integrated into our methodology. As the reviewer notes, these are diagnostics used operationally by centers like ECMWF to predict unusual events against a reference climatology.

However, there is a fundamental difference between these operational metrics and our attribution framework. EFI and SoT are typically applied to evaluate the full forecast ensemble at a given location and time. In contrast, our methodology specifically isolates distributions of extremes using the method of block maxima (i.e., extracting the single maximum or minimum value over a specific time window to fit a GEV distribution). If we were to apply the SoT approach to our distribution of extremes, we would effectively be measuring shifts in the tails of an already-extremal distribution (i.e., the "extremes of extremes"), which would introduce prohibitively large statistical noise. Additionally, EFI does not include forecast values that fall beyond the absolute range of the historical climatology (Zsótér, 2006), a scenario we commonly encounter in our extreme value data (e.g., several model values fall well below the lowest ERA5 value in Figure 5). For these reasons, applying EFI and SoT directly to our extreme value analysis is not statistically suitable.

While it would be possible to analyze EFI and SoT in the SNAPSI experiments using a standard, non-extremal framework, we suggest that doing so falls outside the scope of this study. Our primary focus is strictly on extreme event attribution, employing a probabilistic framework motivated by the broader extreme event attribution literature (as Reviewer 2 notes, Relative Risk and Quantile Shift are standard metrics in this literature). Because EFI and SoT are forecast verification metrics rather than probabilistic attribution tools, we are concerned that adding this analysis would dilute the core aims of the paper. Nevertheless, applying EFI and SoT to the SNAPSI dataset to more generally evaluate forecast skill is a highly valuable suggestion for future investigation.

Minor comments

The control (nudged toward a climatological zonal mean) is useful as a counterfactual, but it may represent an unrealistic state of the system. You already acknowledge this limitation but consider improving the discussion by adding a possible alternative counterfactual construction.

A more realistic alternative might be to nudge each ensemble member towards a randomly-selected observed non-SSW period from the same time of year (as was also suggested by reviewer 2). We have added this suggestion to the text (L497-498).

Figure 2 caption: maximima → maxima. Also, could you use the same colobar for all the plots? The scale is quite similar.

This typo has ben fixed. The three subplots of Fig 2 have the same colorbar, but panel (a) has a downward arrow on the lower range to indicate that some grid points have values falling below this lower range. We have added this information to the caption of Fig 2.

Figure 3: Could you use one colorbar for the plots in the same row? In addition, some colorbars in the left/right column have an upper/lower triangle that it is not present in the equivalent colobar for the other column. Could the top titles have a larger font size?

This figure has been altered in response to Reviewer 2 so that the panels (b), showing IFS statistics, show differences with respect to the ERA5 climatology rather than absolute values. This change more clearly highlights that the differences are small. The colorbars have been made consistent and the font size of the titles increased. As in Fig 3, the arrows on some colorbars indicate that the range of plotted values extends beyond the range.

Line 197:Could you clarify the expression "$(\cdot)_+ := \max(\cdot, 0)$ " in plain English for the reader?

We have added an explanation that the subscript + denotes the positive part of the expression (L199), in addition to this mathematical definition. We will also use 'x' instead of '·' in the definition for clarity.

Line 257: minumum → minimum

This has been fixed (L286).

Figure 9 caption: Non-parameteric → Non-parametric

This has been fixed.

Line 404: intialization → initialization

This has been fixed (L436).

The CNRM model appears in the table as CNRM-CM6-1 and in Figure 9 as CNRM-CM61. Please make it consistent

This has been made consistent (as CNRM-CM6-1) in Figure 9.

References in this response:

Zsótér, E. (2006). Recent developments in extreme weather forecasting, ECMWF Newsletter No. 107 – Spring 2006, pp. 8–1. doi:10.21957/kl9821hnc7