We would like to thank both referees and Nili Harnik again as their constructive comments and suggestions greatly helped us to improve the manuscript. A brief general discussion about the two major concerns raised by the reviewers was provided in our previous comment (AC1, 9th Feb 2022). In the following, we present the specific changes that we have now implemented, accordingly. Our response is divided in three parts: changes based on major concerns, changes based on minor concerns, additional changes. The referees' comments are in *italic gray,* our comments are in blue. Main changes are bold.

Changes Based on the referees' major concerns

1 Description of the methodology and interpretation of causality

Both reviewers raised concerns about the methodology that was applied to quantify the causal relation between polar vortex and subsequent AO extremes.

- We **added subsection 3.5** (Conditional probabilities of polar vortex and AO extremes), where we outline the approach that we follow to quantify the statistical relation between SSWs/ SPVs and AO extremes. Conditioning on stratospheric events allows us to compute the relative probability increase of at least one AO extreme within a subsequent time period. This effectively quantifies the extent to which a stratospheric extreme may act as a predictor of a tropospheric extreme.
  However, quantification of AO extreme occurrence *without* preceding stratospheric extremes, which is needed to assess potential causal links, requires careful evaluation:
  We now make use of the **concepts of attributable risk** (frequently used in other fields such as climate attribution science or epidemiology, see sections 3.5 and 6). In our case the attributable risk **among the exposed** quantifies the fraction of "stratospheric-extreme-exposed" AO extremes that may be attributed to the preceding stratospheric extreme, whereas the attributable risk **among the population** quantifies the fraction of all AO extremes that may be attributed to preceding stratospheric extremes.
- To make the usage of different conditional probabilities clearer, we provide an **overview table** for the applied event definitions (Tab. 2).
- Fig. 7 showed the "estimated probability increase" (AO extremes following SSWs vs. AO climatology). We change the wording to "relative probability increase" to highlight the ratio that appears in the calculation. In addition, we add a plot in the supplement that shows the relative probability increase as a function of the time period that is used to search for AO extremes.
- In section 6, we present composites based on AO extreme events. We compute the probability of negative AO extremes being preceded by SSWs and we had compared this probability to the climatological occurrence of SSWs. We agree with the reviewers that the difference cannot be interpreted as the fraction of AO extremes *being caused* by SSWs (see Fig. 8 of submitted manuscript).
  In the revised manuscript, we **replace Fig. 8** (and Fig. 11., for strong vortex events) with figures that now show that
  - about 50% of AO^{-3} extremes that are preceded by a SSW may statistically be attributed to the preceding SSW, whereas
  - about one quarter of all AO^{-3} extremes may be attributed to preceding SSWs.
- Even though attributable risk is no strict quantification of causality either, it does offer insights into causal links in a statistical sense (see also revised discussion in conclusion section about common drivers).

2 Model Differences

The reviewers noted that given quantitative discrepancies between the models (ECMWF, UKMO) in some of the presented diagnostics, the interpretation needs to be done carefully, and both models could be "wrong". It was suggested that investigating dynamical causes for the observed differences could reveal interesting insights.

- Fig. 6 (relative probability increase of AO extremes following SSWs) shows one of the main results of our study. Considerable discrepancies between the models are observed for negative AO extremes of below –3, as these events occur 40% more often in the ECMWF and 80% more often in the UKMO model, following SSWs. We have now **added the relative probability increase of positive AO extremes following SSWs**, which is negative, i.e., these events become less likely following SSWs. Both models show quantitative agreement for AO thresholds up to about 2.5 standard deviations. To check whether observed differences for thresholds of ±3 stem from sampling uncertainty, we added 95% confidence intervals, obtained via bootstrapping. Indeed, results for AO $\gtrsim \pm 3$ are associated with considerable uncertainties, however, they cannot fully explain the observed differences.
  We believe that potential dynamical sources that can lead to such differences are numerous. They could be related to intrinsic tropospheric dynamics (e.g., related to wave-mean flow feedbacks) or to the manifestation of teleconnections related to external forcings (e.g., from the tropics). Generally, the analysis of extreme events is expected to be very sensitive to even tiny modulations of the underlying distribution. We address the model discrepancies in the discussion.

Changes Based on the referees' minor concerns

Referee 1

Other comments
L18: What does it mean: “up to a degree of 27%”

We have adapted the statement to the new methodology that is used to attribute AO extremes to preceding SSWs. It now says: "3) approximately 50% of extremely negative AO states that follow SSWs may be attributed to the SSW, whereas about one quarter of all extremely negative AO states during winter may be attributed to SSWs."

L31: Please clarify whether you cite daily AO index value, monthly value or seasonally value.

We added "daily" [AO index...].

L34: Do Kim et al discuss wildfires in winter or in another season?

Kim et al. report that wildfires occur predominantly around April and find that the annual total burned area in southeastern Siberia is significantly correlated with the average AO in February-March. We added "in February and March".

L54: “are needed” for what?

We try to make it clearer by adding [Therefore, a very large sample of SSW and SPV events are needed] "to quantify the subsequent risk increase of AO extremes".

L146: Please explain what does “dynamical SSW” mean and provide reference if it has been introduced elsewhere.

We expanded the paragraph to explain more detailed how we defined dynamical SSWs as the logical intersection between SSW and SSD (Sudden Stratospheric Deceleration) events.

L151: “we therefore do” what?

Thank you for noting, we now write: "[...] we therefore focus on SSWs only, to allow better comparison with other studies."

Figure 1: Although interannual variability of predicted SSW frequency is not the main point of your article I wonder if upper panel of Fig. 1 could show relative frequency of p- SSW rather than absolute numbers. It is quite exciting to see so small number of p-SSWs in 2008/09, a winter in which an SSW occurred in the real world.

We appreciate the idea and agree that the relative frequency provides interesting additional information. As we are not aware of a standard procedure to derive seasonal SSW probability from ensemble forecast data, we introduced a **proxy for the seasonal SSW probability**, that is described in detail in **appendix B**. In Fig. 1, we added the proxy as a subplot.

L165: “the event was generally very rare” sounds strange to me

We now write that the strong vortex lead to "an only marginal SSW probability in the forecasts", which suggests that "the event itself was unlikely given the prevailing dynamics".

L175: Please provide equation which you apply

We decided to only mention the number in the text and refer to the new appendix B. The calculation follows the same procedure as for the newly introduced seasonal SSW proxy.

L197: A rather complicated deseasonalization approach has been used. Why not used a simpler approach in which climatology is estimated using other hindcast years? For example, for ECMWF hindcasts this would provide 19x11=209 realization to build a climatology for each date and lead time. Why do you think it is not enough?

With our approach, we follow the procedure described here: https://www.ecmwf.int/en/forecasts/documentation-and-support/extended-range/re-forecast-medium-and-extended-forecast-range. As our analyses focus on extreme events, we particularly require an accurate representation of the distributions' tails, which we ensure by including forecasts from within a plus/minus 2 week window. We added a comment in the conclusions section.

L219: “occur only few days after the event” can you provide the exact lag?

We added "(lag +2 days: $-6 ms^{-1}$)".

L223: I do not think NAM1000 distribution is significantly different from 0 at negative lags.

We have tested whether the described positive anomalies are significantly different from 0 via a two-sided student's t-test and find p-values below 10^{-10} in both ECMWF and UKMO models (as a result of the large sample size). However, we agree that physical interpretation requires further research.

L226: the trend goes to weaker negative values, not positive.

Corrected, thank you.

L234: I am not sure the name “ECMWF S2S model” is correct.

We now write "ECMWF model".

L236: “most phases of negative NAM1000”, perhaps: “most cases of negative NAM1000”

Adapted as suggested, thank you.

L243: I do not think NAM1000 in ERA5 follows AR1 process either, or have you checked it?

Based on the concerns raised, we have checked again and identified a bug in the previous calculation for AO persistency in ERA5 (thank you!). It turns out that ERA5 and ECMWF (S2S) agree very well in their climatologies, suggesting that the observed variability cannot be reproduced by an AR1 process. We have updated the text accordingly.

L258: Should not probability of negative NAM be exactly 50%, by construction?

The NAM index distribution turns out to be not perfectly Gaussian, therefore, mean and median are not exactly equal. We now write: "Asymmetry between positive and negative values arises from the AO distribution that is not perfectly Gaussian (skewness: -0.13)."

Figure 6: What is the period used for calculating the probability increases?

We added that the we compute the relative probability increase by averaging over periods of 25 days to 40 days.

L429: I do not think that increasing number of models would help to make definitive quantitative statements unless you know which models are right and which models are wrong. Since all models are different you could only possibly increase the spread.

Yes, we agree and we removed the sentence. We provide a brief discussion about observed discrepancies between ECMWF and UKMO. Including further models would likely raise similar issues.

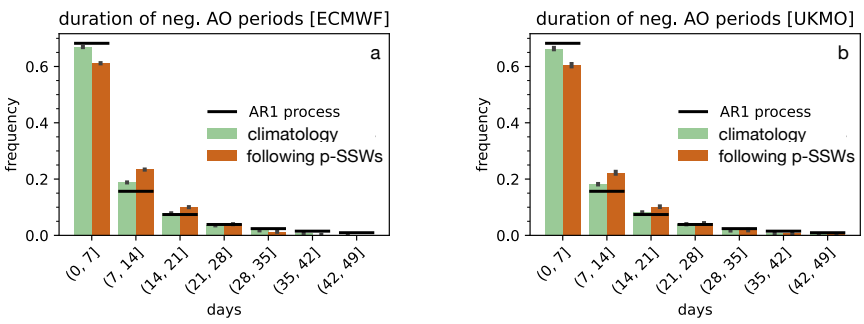
Referee 2

L143: Will be the results sensitive to the WMO's definition that includes the reversal of the meridional temperature gradient?

We have chosen to define p-SSWs and p-SPVs based on u60 alone, mainly to follow the standard definition of SSWs that is used most often in the literature and to limit the required amount of data storage. Even though we have not explicitly tested, we would expect modest differences in the classification of individual events, but we would expect that differences average out in the composite mean (see, e.g., Butler et al., 2015). Furthermore, our analysis of dynamical SSWs can serve as a sensitivity analysis, and the results show only minor quantitative differences.

Figure 3. Please also add a similar histogram for UKMO model next to this figure. Also please add the uncertainty in this plot.

Both histograms for ECMWF and UKMO agree extremely well. UKMO forecasts show very long periods (>28 days) of persistently negative AO slightly more often, which is likely due to the longer leadtime. In addition, sampling uncertainties are very small, such that error bars denoting the 95% confidence intervals are even hard to see. Therefore, we decided to omit error bars as well as the histogram for UKMO data, but we added a comment.



L175: Please delete this and just mention the number. Otherwise, please provide a full equation before inserting the number.

We now provide only the number and refer to the new appendix B for further details.

Figure 4. Do you have a similar figure for ERA5? How does it look like compared to UKMO and ECMWF models? It's hard to get definitive quantitative statements since both the model are probably not right.

We have checked: AO<0 is more likely in ERA5 at positive lags (between 55% and 75%) and even at negative lags already (up to 60%). However, 95% confidence intervals span up to 40% and therefore reveal large associated sampling uncertainties. Neither AO<–3 nor AO>+3 events are observed following SSWs in ERA5, likely due to the limited sample size. We added a comment.

L429: I dont think you will get a definite answer for this rather than a spread of the quantification of the probability of extreme AO events following extreme strat. events in different model configuration.

Yes, we agree and we removed the sentence. We provide a brief discussion about observed discrepancies between ECMWF and UKMO. Including further models would likely raise similar issues. [see response to referee 1]

Additional Changes

- To avoid confusion and to allow for shorter abbreviations for event definitions, we changed the terminology from NAM1000 (Northern Annular Mode at 1000hPa) to AO (Arctic Oscillation).
- We use the notation AO^- for negative and AO^+ for positive AO events. Particular thresholds are explicitly indicated, e.g., AO^{-3}.
- We adjusted the way we compute the probabilities for "at least one event within time t", e.g. $P(AO_{wt} \mid SSW)$. Before, we had computed daily event probabilities and derived "at least one" by: 1 minus "no event on every day". Now, we compute the probabilities by explicitly checking how many of the available forecasts fulfill the respective conditions. Both approaches are expected to yield the same result in the limit of large forecast sample sizes.