We would like to thank both referees for carefully reading our manuscript and we appreciate the constructive feedback. In addition, we would like to thank the editor, Nili Harnik.
Two important issues that both referees highlighted to us are the 1) differences in the results between the two models, and 2) the methodology, particularly with regard to interpretation in terms of causality. Therefore, our response below begins with a general discussion regarding these two issues. Thereafter, we address the specific concerns and suggestions raised by the reviewers and outline how we plan to adapt the manuscript, accordingly.
The referees' comments are in *italic gray,* **our responses are in** blue**.**

Model Differences

We agree that the differences between the two models are interesting and that they need to be considered for the interpretation of the results. However, we note the following points that may help to put the observed differences into perspective.

Generally, the analysis of extreme events is expected to be highly sensitive to (tiny) modulations of the underlying distribution. Our results show that both models agree well in terms of the *average* AO shift following stratospheric events (Figs. 2, 9). Differences by a factor of about 2 are observed for the analysis of extreme events (e.g., Fig. 6). However, we believe that these discrepancies are not overly severe, since small changes in the distributions' tails could translate quickly into different orders of magnitude.

Furthermore, the differences between the models come out most prominent in Figs. 6 and 10, where the relative risk increase of AO–/ AO+ is presented following SSWs/ SPVs. However, we note that particularly the relative risk increase is very sensitive to changes in the climatological occurrence of AO extremes, which appears as a small number in the denominator. With this in mind, we plan to include the *absolute risk difference*, which may complement the analysis.

In discussing potential physical processes that could explain the observed model differences, we will consider, e.g., the reference noted to us by the editor. However, a detailed analysis of the model differences is beyond the scope of our study.

Methodology and causality

We appreciate the reviewers' valuable feedback, which has helped us to clarify a few important points. We agree that the manuscript will benefit from an improved description and a more accurate interpretation of the methodology. Therefore, we plan to take the following steps:

I. We will provide an overview table that summarizes our event definitions and shows how conditional probabilities are derived from the forecasts.
II. Based on the conditional probabilities that are stated in the table, we will provide equations for the diagnostics displayed in Figs. 5, 6, 7, 8, 10, 11. Where appropriate, we will adapt the terminology to be consistent with language that is used for similar problems in other fields (e.g., "relative risk increase" and "(absolute) risk difference" in medical statistics).
III. We recognize that the interpretation of our questions 2 and 3 need to be carefully separated in terms of causality.
   ad Q2: We found that stratospheric events enhance the probability for AO extremes relative to climatology, therefore, the stratospheric event can be considered a (statistical) predictor. As noted by the referees, our analysis is not proof of causality in a strict sense because other processes could be involved. We are currently in the process of analyzing the role of such processes, for instance, SST variability associated with ENSO and plan to include relevant conclusions from this analysis in the discussion. However, a detailed analysis of further processes is beyond the scope of the present study and is left for future work.
   ad Q3: We have analyzed how often AO extremes are preceded by certain stratospheric events and compared the number of observed events to the number of expected events based on climatology. We recognize that this analysis is not able to prove causality, even in the absence of other processes. While we think that a quantification of the exceedance

occurrence of stratospheric events ahead of AO extremes is still useful, it requires a fundamentally different interpretation compared to our analysis of Q2. Instead of "caused by p-SSW"/ "caused by u60<0" (e.g., Figs. 8, 11) we will refer to the exceedance probability of preceding p-SSWs/ u60<0.

Response to comment by anonymous referee #1

The article by Jonas Spaeth and Thomas Birner entitled "Stratospheric Modulation of Arctic Oscillation Extremes as Represented by Extended-Range Ensemble Forecasts" discusses the influence of the stratospheric polar vortex on the surface climate. This topic has been discussed a lot in the literature and it is well established that extreme states of the polar vortex tend to shift Arctic oscillation towards certain states with implications for regional weather conditions. However, the strength of the link between stratospheric conditions and surface weather is poorly estimated because of a low signal-to-noise ratio. To address the low signal-to-noise ratio problem the authors turned their look towards model simulations which provide much more data sufficient to obtain robust estimates of the stratosphere-troposphere coupling. The underlying assumption is that the models provide a reasonable representation of the real atmosphere. The assumption seems to be violated at least in some cases because some estimates obtained from the two different models considered (ECMWF and UKMO) diverge significantly. Which of the two models is closer to the real world is difficult to establish. Therefore, the interpretation of the results should be done carefully. Nevertheless, the article presents novel results which in my opinion go beyond state-of-art. I believe the article can be published in Weather and Climate Dynamics after revision.
My main criticism concerns the causal analysis. This should be better described and, in case if the approach used by the authors is a well-established one (I apologize for my ignorance), proper references to background literature should be made. Additionally, I strongly recommend a language check before publication.
Major comment:
The authors distinguish between increased probability of AO extremes following stratospheric events (their question 2) and how often stratosphere can be considered a cause of extreme AO events (question 3). Both questions are addressed in terms of probabilities (e.g. Figs. 6,7,8). While I understand the difference between the two questions in principle, I do not understand how you manage to solve them separately, given that both questions can only be answered in statistical sense. In figure 7 you show probability of at least one SSW day preceding a randomly sampled day; however in Figure 8 this probability becomes a probability of AO extreme preceded by a SSW day by chance (left panel of Figure 8). This is not the same, clearly. Assume hypothetical world in which all AO extremes are caused by an SSW occurred during previous 30 days. Then dashed lines in Fig. 7 would reach 1 by day -30. However, this would not affect your climatology because it only measures probability of SSW. As a result, you would never be able to correctly answer question 3 using your methodology. For day 30 you would only obtain the difference between 1 and an SSW probability, which is not the right answer to question 3.

We agree that our method underestimates the true causal relationship, because part of it is masked by the climatological occurrence of SSWs. The effect will be larger if the true causal relation is strong (i.e., almost all AO extremes caused by SSWs) and if the SSWs occur very often in the climatology. What our method hence quantifies the exceedance probability (above climatology) of preceding SSWs to AO– extremes. Nevertheless, this does not represent a rigorous quantification of the causal relation, even under the absence of other processes. We thank the referee for this comment and adapt the manuscript accordingly (see our initial comment for more details).

Figure 11 illustrates the same problem – there is no evidence in data that AO>3.5 can occur without an SPV within previous 40 days, yet only about half of those events that occurred

after SPV can be attributed to SPV following your methodology. I believe the methodology needs to be revised (or I miss something).

Same issue, the nomenclature and interpretation will be adjusted accordingly.

We reply to a few selected minor comments below. All remaining comments will be addressed in the revised manuscript.

Other comments
L18: What does it mean: "up to a degree of 27%"
L31: Please clarify whether you cite daily AO index value, monthly value or seasonally value.
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 will add "in February and March".

L54: "are needed" for what?
L146: Please explain what does "dynamical SSW" mean and provide reference if it has been introduced elsewhere.

Instead of the commonly used definition of SSWs, which is based on the reversal of u60 from westerly to easterly, e.g., de la Camara et al. (2019) analyze sudden stratospheric deceleration (SSD) events, in order to better account for rapid changes in the dynamics. Our definition of dynamical SSWs forms the intersection of SSWs (accounting, e.g., for fundamentally different wave propagation properties in easterly winds) and SSDs (ensuring a rapid deceleration around the SSW central date). Our results reveal only modest quantitative differences between SSWs and dynamical SSWs, we therefore focus on SSWs only, to allow better comparison with other studies. We will make the description more detailed.

L151: "we therefore do" what?
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 had decided to show the absolute numbers of p-SSWs so that the contribution of individual years to the overall analysis can be read off. We agree that the relative frequency would also be interesting and we will include a comment about the interannual variability of SSW frequency per winter.

L165: "the event was generally very rare" sounds strange to me
L175: Please provide equation which you apply
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. However, this comes at the cost of not accounting for potential leadtime dependent model biases.

L219: "occur only few days after the event" can you provide the exact lag?
L223: I do not think NAM1000 distribution is significantly different from 0 at negative lags.
L226: the trend goes to weaker negative values, not positive.
L234: I am not sure the name "ECMWF S2S model" is correct.
L236: "most phases of negative NAM1000", perhaps: "most cases of negative NAM1000"
L243: I do not think NAM1000 in ERA5 follows AR1 process either, or have you checked it?
L258: Should not probability of negative NAM be exactly 50%, by construction?

The NAM index does not follow a perfect Gaussian distribution, therefore, mean and median are not exactly equal.

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

We agree and we will describe that including additional models may help to better estimate the robustness of the results.

Response to comment by anonymous referee #2

The authors used a large set of extended-range ensemble forecasts within the sub seasonal-to-seasonal (S2S) framework (namely ECWMF and UKMO models) to obtain an improved characterization of the modulation of AO extremes due to stratosphere- troposphere coupling. Within this framework, they investigated how much stratospheric polar vortex extremes increase the probability of persistently AO phases and their extremes. They found that following potential SSW events, persistently negative AO states (> 1 week duration) are 16% more likely, and the likelihood for extremely negative AO states (< –3σ) is enhanced by at least 35%. How the stratospheric polar vortex extremes can be considered as the cause of the subsequent AO extremes was also quantified and discussed. Despite the straightforward analysis presented in this paper, I still found the results are interesting and the diagnostics can be useful for the forecast model assessment. The main issue I have is a lack of dynamical analysis to explain the differences in the two models in representing the AO extremes followed by stratospheric events (SSWs and SPVs) and the results regarding causal relationships between AO extreme and stratospheric polar vortex extremes. Hence my suggestion is major revisions. Once my points below are answered, I can recommend this work to be published in WCD.

General Comments:
Two S2S forecast models (ECMWF and UKMO) used in this study showed some quantitative disagreement (i.e., the results diverge significantly e.g., Figs. 5, 6, 7. 8 etc). However, there is no dynamical analysis and explanations to address the issue rather than simply comparing the results in a statistical sense. It would make the results clearer if you could address this issue in the paper.

We agree that the disagreement between the two models are interesting and that the analysis would benefit from investigating the underlying dynamical causes. We address the observed differences between the models at the beginning of this response.
Beyond that, we think that potential causes are numerous (for instance, differences in wave-mean flow feedbacks or external forcings, e.g., from the tropics). We are in the process of investigating potential sources to the observed differences and we will include the outcome in the discussion section. However, a detailed analysis would go beyond the scope of this study.

I am not convinced about the causal analysis in this paper. As you are aware, the extreme AO events are not only proceeded by extreme stratospheric events, but also by mid-latitude

winter circulation such sea-level pressure, sea-ice and remote forcing from the tropics. How can you isolate the possible stratospheric influence alone from these other factors (since this may not direct/linear statistical relationship)? I believe that not all stratospheric polar vortex extremes lead to AO extremes. You probably need to revise your methodology to address this question.

We agree that statements about causality in a strict sense are not possible with the kind of analysis presented here. Based on the reviewer's comments we have decided to replace our terminology with more neutral statistical terms such as predictors, probability exceedance and the like. We acknowledge that multiple predictors might exist and not all predictors are based on direct causal relations.

That said, note that other tropospheric drivers of AO events only represent an issue for our statistical inference of stratospheric influence if they have an impact on the stratosphere themselves. Nevertheless, a modification of the stratosphere by such a third driver (e.g., Ural blocking, which may influence the AO directly) might further lead to modifications of the AO. In this sense individual drivers are generally difficult to isolate (because they tend to be coupled), even if their causal connection is clear in principle. Again, we will modify wording to speak of "predictor" in such cases, which should circumvent the issue of common coupled drivers.

As one example of a common quasi-external driver we are in the process of evaluating potential influences of different ENSO states and will include some aspects of these results in our discussion (although a more in-depth study in this direction seems of merit and will be left for future work).

We reply to a few selected minor comments below. All remaining comments will be addressed in the revised manuscript.

Other Comments:
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.
L175: Please delete this and just mention the number. Otherwise, please provide a full equation before inserting the number.
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 expect ERA5 to be extremely noisy in this diagnostic due to the limited sample size in observational record, but we will check.

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.

We agree, the statement will be modified (see response to RC1).

Response to comment by editor, Nili Harnik

Both reviewers raise concerns about the interpretation in terms of causality, which should be addressed.
In addition, both reviewers comment on the need to carefully interpret the results given the large differences between the two models used. In this regard, I am wondering if the following paper, shows biases in the eddy energy spectra in ECMWF IFS model., has anything to do with this difference - Augier and Lindborg, 2013: A New Formulation of the Spectral Energy Budget of the Atmosphere, with Application to Two High-Resolution General Circulation Models. JAS, 2293-2308.

Thank you, we appreciate the reference to the study. As describe, we are currently investigating which processes could lead to differences in the results and we will consider potential differences in the eddy energy spectra.