Response to Reviewers - 1st revision

Dear Editor,

We would like to thank all reviewers for their reviews of our manuscript and their insightful comments. Please find our detailed responses to the reviewers' comments and suggestions below. The changes have been included into the manuscript (indicated in **bold**). All line numbers refer to the new (annotated) version of the manuscript.

Sincerely,

The authors

1 Reviewer 1:

Major comments

1. Had you considered to split the 28 days period into, for example, two parts (week 1-2 and week 3-4) in the first part of the manuscript? It would probably be interesting to look at least at U850 and cyclone frequency anomalies especially having in mind the results of the predictability part.

We thank the reviewer for this comment. We have revised the manuscript and added a two new figures (Figure 10 and Figure 11) to demonstrate the time evolution of cyclone frequency response following extreme stratospheric events (SSWs and strong vortex events). Given the results of the predictability analysis (Figures 7-9), we have decided to focus on the differences between the two stratospheric conditions (weak/strong vortex). While these events are generally considered to have similar but opposite tropospheric response, there are also differences between them, especially from the predictability aspect.

In the new Figure 10, we plot the time evolution of cyclone frequency anomalies (ensemble mean) in the North Atlantic Basin (see box in Figure 3) in the period following SSW and strong vortex events, and compare between successful and unsuccessful predictions. We also analyze the MSLP and Z'100 anomalies corresponding to these periods. We find that capturing the lower-stratospheric circulation after SSW events is a necessary but not sufficient condition for predicting the downward response (i.e., the greatest difference between successful and unsuccessful prediction is in the troposphere), whereas for strong vortex events capturing both the lower-stratosphere and the troposphere states is necessary.

In Figure 11, we evaluate the uncertainty in the prediction of cyclone frequency after these events, based on the average ensemble spread (the ensemble mean is superimposed in black contours). This additional analysis provides insight regarding the large latitudinal differences in the ensemble spread, suggesting a limited capability of the

reforecasts in reproducing the response in mid- and high-latitudes.

2. The box position choice does not seem well explained. You say that in this region the increase in cyclone frequency is biggest after SSW events (L185), but the anomalies are biggest only in reforecasts (Fig. 3a). Moreover, the anomalies are biggest and statistically significant over the Northern Europe in reanalysis (Fig. 3c), which can be also seen in reforecasts. Had you considered taking a box more to the north-east of its current position? Also, was your choice of the box position based only on the anomalies after the SSW events? I see that the biggest anomalies after the SPV cases are still concentrated inside the box (Fig. 3b), but maybe this can be pointed out in the text.

The boundaries of the box were determined according to the region of largest increase in cyclone frequency after SSW events (considering all 14 events in the reanalysis dataset). We agree with the reviewer that a larger or shifted box can better capture the anomalies after both SSWs and SPV events. Therefore, we have shifted the southern boundary to 35°N and extended the northern boundary to 55°N. In the revised version, we focus our analysis on the mid-latitude region (35°-55°N) of the North Atlantic (60°W-0°E). This region, located on the southern flank of the North Atlantic storm track, is where the change in cyclone frequency after SSW and strong polar vortex events is the largest. We have updated all the plots in the manuscript according to the new box definition.

Minor comments

L32 "predication" -¿ prediction
Corrected.

L34 Add brackets to citation
Corrected.

L48-51 This sentence seems to repeat the information given above, please consider removing it or rephrasing the repetition
We have rephrased that paragraph as suggested by the reviewer to avoid repetition.

L88 As reforecasts are initialized in conjunction with real-time forecasts, could you provide here the dates/years of the real-time forecasts? You indicate below the model versions used, but this is potentially confusing, as, for example, the 46R1 version does not have reforecasts for December

Dataset/Forecast	Operation period
Cycle 46R1	from 11/06/2019
Cycle 47R1	from 30/06/2020
Cycle 47R2	from 11/05/2021
Cycle 47R3	from 13/10/2021

Table 1: Implementation dates of each ECMWF model version. Source: `https://confluence.ecmwf.int/display/S2S/ECMWF+Model`.

2019 2.
The ensemble re-forecasts consist of a 11-member ensemble starting the same day and month as a real-time forecast (Monday and Thursday), covering the past 20 years. The implementation dates of each ECMWF model version is summarized in the table below (Table 1). For example, the reforecasts of December 2, 2019 belongs to the model cycle CY46R1 since it has been computed between 11/06/2019 and 30/06/2020. The reforecast for this date has been initialized on same date as the real-time forecast of 02/01/2020. This reforecast consisted of a 11-member ensemble starting on 2nd January 2000, 2nd January 2001,... to 2nd January 2019 (20 years).
To clarify this point, we have added this information to the Methods section (lines 80-85).

L97 Please consider adding "... in the ECMWF model and in reanalysis..." if you used the same algorithm
Corrected.

L100 Consider adding here a remark that the number of the cyclone tracks can be found in Fig.6
Corrected.

L104 DJF -¿ DJFM for consistency throughout the text
Corrected.

L108 Did you use cross validation when computing the anomalies for each ensemble member?
Cyclone frequency anomaly for each ensemble member is computed as the difference in the number of cyclones detected in the 28 days after the SSW and the climatological cyclone frequency for this period. As the computation of these anomalies is mathematically straightforward (i.e., an anomaly is defined as a deviation from the climatological mean), we have not performed cross validation when computing the anomalies.

L108 While I understand the choice of 28 days, it could be better clarified here for better understanding

We added a clarification in the text regrading the choice of the 28 days period, as follows (lines 110-115):

In the NH, anomalies in the tropospheric circulation after extreme stratospheric events (such as SSWs, weak vortex events and strong vortex events) can persist for up to 60 days after their onset (Baldwin and Dunkerton, 2001), and thus may prove to be useful for tropospheric weather prediction. A period of 28 days after the onset of SSWs and strong vortex events is chosen in order to understand the initial tropospheric response and its potential for subseasonal predictions of the surface response.

L117 I wonder of you checked if there is no difference indeed when using ERA-Interim or ERA-5?

We have replaced the analysis using ERA-Interim (detection of SSW and SPV events) with ERA-5 data instead. We have updated the manuscript such that the entire analysis is performed using ERA-5 reanalysis. The change does not affect the results of the paper.

L120 Please specify that the list given in (Butler and Domeisen, 2021) contains only final warming events, rather than all warming events. Or consider omitting this part of the sentence

We rephrased this part of the sentence to clarify this point.

L131 Please consider mentioning that the dates of the SSW and SPV can be found later in Figures 7 and 8. I wanted to suggest adding a table with dates, but it seems excessive, since the information appears later in the text.

A full list of SSW and SPV dates that are used in this study can be found in Figure 7 (and Figure 8). We added this information to the text (line 134).

Fig 1c The model bias spans from -4.5 to 4.5% while the frequency itself changes from 0 to 45%, do you think that the bias is statistically significant in this case?

Generally, we find the model bias in cyclone frequency to be statistically significant compared to the reanalysis in most of the North Atlantic sector. In the revised version, we have added an analysis of the statistical significance of the cyclone frequency bias compared to reanalysis in Fig. 1, panels c-d. In these panels, we show that climatological cyclone frequency bias, for model initializations in November to March. We find that the bias is relatively small when considering the first 7-days after the initialization, compared to larger, more statistically significant biases found in the 28-day average. Thus, biases in the range of -4.5 to 4.5% are significant, when evaluated

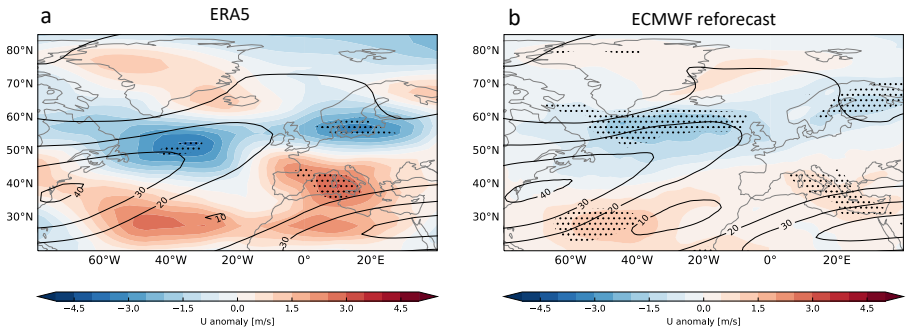

Figure 1: Zonal wind anomalies following SSW events in (a) the ERA5 reanalysis, and (b) the ECMWF reforecasts, averaged over 13 SSW events between 2000 and 2019, excluding February 2010. The black contours show DJFM climatology.

against climatology.

L154-155 Repetition of "in ERA-5"in the sentence, please remove one of them
Corrected.

L155 Did you look at the individual events before constructing the composite? It would be interesting to know which events had stronger response. However, the washed-out signal in the reforecasts might show that the model underestimates the response, especially averaged over 28 days of forecast.
Yes, we have examined individual events before constructing the zonal wind composite. We found that some specific events, such as the February 2010 SSW event, exhibit a particularly strong signal in the zonal wind response in the observation. Excluding this event, for example, shows that despite a similar magnitude of zonal wind anomalies, the model underestimates the response in the 28-day average of the forecasts (as shown in Figure 1 below). This result is consistent with the results in the manuscript. Additional information on the individual response of SSW and strong vortex events can be found in the revised manuscript (section 3.5, Figures 7-9), showing the ensemble response for the North Atlantic (Figure 7) and Europe (Figure 8).

L196 "The statistical significance of this shift. . . " it is not clear whether you refer to the shift compared to all winter days, or the small shift of reforecasts compared to reanalysis
We refer to statistical significance of the shift in the distribution of cyclone frequency anomalies compared to all winter days. We rephrased this paragraph (lines 205) to clarify this point.

Fig. 4 As I understand, the figure shows cyclone frequency anomalies after the 14 events in each subplot, but in this case what does the height of the bars show? Counts on y-axis does not add up. If you used a somehow broader statistics, please clarify that in the caption.

In Figure 4, we analyze the distribution of the cyclone frequency anomaly for each type of stratospheric event (SSW or strong polar vortex event). Anomalies are averaged over a period of 28 days (days 1-28 with respect to the central date of the stratospheric event). Since we have 14 events for each event type, the total number of counts is [N=14] for ERA5 reanalysis and [N=140] for the reforecasts (14 events x 10 ensemble members]). The left y-axis in Figure 4 shows the probability density, hence each bin displays the bin's count, divided by the total number of counts and the bin width, so that the area under the histogram integrates to 1.
To address the reviewer's comment, and to provide a clear comparison between "probability density" and "raw counts", a histogram of cyclone frequency anomalies following SSW and strong polar vortex events in terms of raw counts for both reanalysis (panels a,b) and reforecasts (c,d) is shown below (Figure 1). Displaying the probability density allows a direct comparison between the distribution of anomalies in the reanalysis and in the reforecasts (purple bars). Therefore, we have decided to remain with the "probability density". We have clarified this information in Figure 4's caption, and changed the y-axis label from "counts" to "probability density".

Fig.5 and L212 Could you explain why there are more cyclone tracks (black lines) detected in reforecasts than in reanalysis? I suppose that you used each ensemble member separately rather than ensemble mean, which could be mentioned in text for easier understanding. Also, you mention in Data and Methods that in this part you use more reforecasts from three model versions, but could you explain more in detail why do you use other model versions. The temporal resolution increase to 6-hourly data is understandable here.

We have added this information to the text to address these important points (lines 212-215). There are more tracks in reforecasts than in reanalysis due to the use of all available ensemble members (11 members) rather than the ensemble mean.
Regarding the use of three different model version, this is a result of implementation dates for each new prediction system version (cycle) by the ECMWF. We have included the table with the specific dates (Table 1 in the response letter). Depending on the time in which data has been downloaded, the model version of the reforecast will be different.

L248 Did you check this correspondence case-by-case here, rather than the overall ratio?

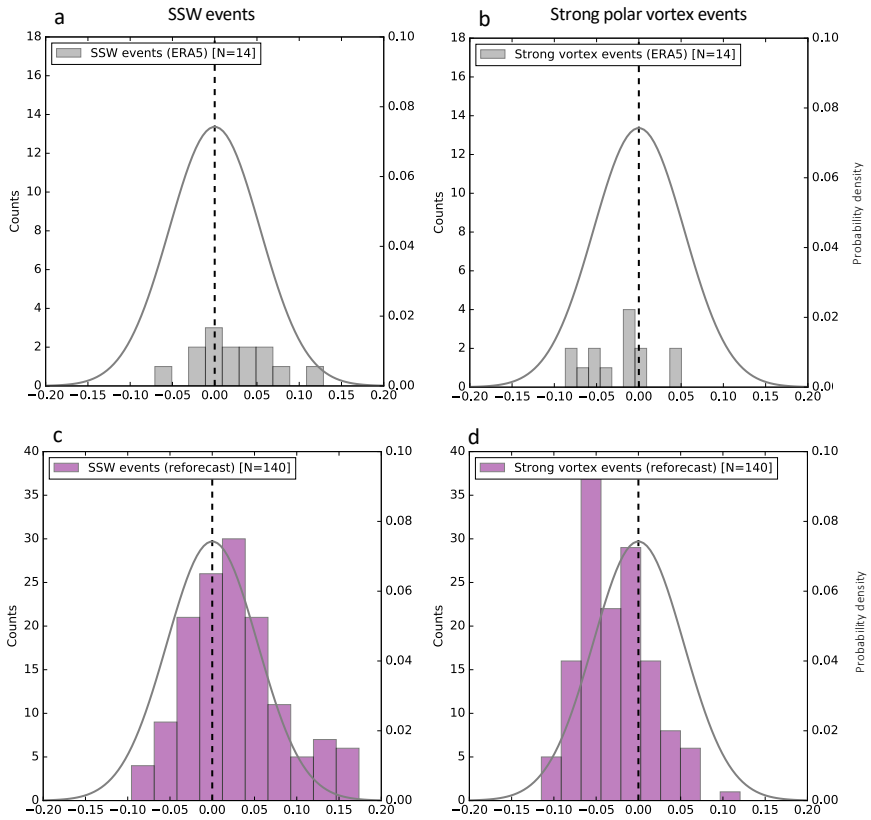

Figure 2: Histogram of cyclone frequency anomalies (in counts) following (a,c) SSW and (b,d) strong polar vortex events in ERA5 reanalysis (grey) and in the ECMWF reforecasts (purple). Anomalies are averaged over the mid-latitude North Atlantic. The grey curve in each panel indicates the climatological probability density for all days in DJFM in the reforecasts.

The ratio of ensemble members with and without a 'canonical' downward response is based on case-by-case examination of the response, as shown in Figures 7-9. In the revised manuscript, this ratio has slightly changed from the previous manuscript due to the new box position in Figure 3. However, we would like to emphasize here that the focus of this manuscript is general behaviour of the downward response in periods following stratospheric events and how well the response is predicted by the ECMWF prediction system. Thus, more emphasis is given to the composite analysis in the first part of the paper. We can easily include the case-to-case analysis of zonal wind and cyclone frequency in the Appendix, if the reviewers would find it useful.

Fig.7b,d It would probably be better for understanding if you indicated in the figure that N=10 for enhanced cyclone frequency and N=4 for the reduced, etc.
We have now added this information to the plot (see x-axis labels).

L261 Did you have a look why the week-1 hit rate for 11 Feb 2005 was so low, especially considering that the skill is higher on the following weeks and the averaged skill is rather high (0.7 from Fig. 7c)?
The reasons for the change in skill for the 11 Feb 2005 SSW events are yet unresolved. Analyzing the potential causes requires a more focused study on the dynamics of this specific event, as was done, for instance, for the 2018 SSW event (e.g., Karpechko et al., 2018, Kautz et al., 2020). While this is an interesting question, we believe this analysis is out of the scope of this paper.

L270 "... predicted a weakening of the cyclone frequency in the period that followed the SSW." As I understand the majority of ensemble members still predicted the increased cyclone frequency on week 1 and 2 in this case, so maybe you can specify that it is not about the period that directly follows the SSW.
Thanks for this comment, we rephrased this sentence to specify that we refer to week 1 and 2 in this case, and not for the entire period.

L297 It could be worth specifying that in case of SSW it is about the reduced frequency
We have rephrased that sentence to clarify that.

L308 Consider adding "after SSW events in these cases." as temperatures are not always predicted poorly after SSWs.
We corrected that.

2 Reviewer 2

2.1 Major comments

1) First, the study examines only one subseasonal model (ECMWF) and therefore lacks a generalized view of how other leading subseasonal prediction systems reproduce the stratosphere-North Atlantic storm track relationship. The fields from the reforecasts of the other models are readily accessible and possible to be analyzed and compared/contrasted. I am not necessarily advocating using every model, but I think adding a few more will be very useful and strengthen the message.

We thank the reviewer for the comments. Indeed, we agree that a systematic analysis of model biases in the downward impact of extreme stratospheric events across a wide range of subseasonal forecast systems would be an important step towards a better understanding of the role of the stratosphere for prediction of surface climate on subseasonal to seasonal timescales.
However, inter-comparison studies of S2S prediction are way more complex, and require more effort (computationally, as well as time-wise), more data, and usually are done as large community/collaborative studies, as was recently done, for instance, in the Lawrence et al., 2022 for a systematic analysis of model biases in the stratosphere. In the context of extratropical cyclone analysis, this would require implementation of the cyclone detection algorithm for identification and tracking of extratropical cyclones across all prediction systems available in the S2S project.
One of the reasons for using the ECMWF prediction system in this study is due to its more highly resolved stratosphere (relative to other models in the S2S project, e.g., Domeisen et al., 2020). Overall, the ECMWF model has been shown to have a good representation of the variability in the stratospheric polar vortex, in terms of extreme event magnitude and the associated dynamical drivers (Wu et al., 2022). Furthermore, the model represents mid-latitude storm track well, as shown in Fig. 1 in the revised manuscript.
Performing a more complex, extended study on the S2S biases in the prediction of the storm track in the future, may be considered as a follow up on the current manuscript and its main findings using the ECMWF extended-range prediction system.

References:

- Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., ... and Taguchi, M. (2020). The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere-troposphere coupling. Journal of Geophysical Research: Atmospheres, 125(2), e2019JD030923. https://doi.org/10.1029/2019JD030923

- Lawrence, Z. D., Abalos, M., Ayarzagüena, B., Barriopedro, D., Butler, A. H., Calvo, N., de la Cámara, A., Charlton-Perez, A., Domeisen, D. I. V., Dunn-Sigouin, E., García-Serrano, J., Garfinkel, C. I., Hindley, N. P., Jia, L., Jucker, M., Karpechko, A. Y., Kim, H., Lang, A. L., Lee, S. H., Lin, P., Osman, M., Palmeiro, F. M., Perlwitz, J., Polichtchouk, I., Richter, J. H., Schwartz, C., Son, S.-W., Statnaia, I., Taguchi, M., Tyrrell, N. L., Wright, C. J., and Wu, R. W.-Y.: Quantifying stratospheric biases and identifying their potential sources in subseasonal forecast systems, Weather Clim. Dynam., 3, 977–1001, 2022. https://doi.org/10.5194/wcd-3-977-2022

- Wu, R. W. Y., Wu, Z., and Domeisen, D. I. (2022). Differences in the sub-seasonal predictability of extreme stratospheric events. Weather and Climate Dynamics, 3(3), 755-776. https://doi.org/10.5194/wcd-3-755-2022

2) Next, I found aspects of the methodology confusing. In the methods section, the authors mention that they use the ERA-Interim reanalysis product for determining the state of the polar vortex (strong vs weak) but then use ERA5 for their analyses. Determination of events in the ERA5 dataset is very straightforward. So, to be consistent, the authors should use one reanalysis only throughout their work.

We agree with the reviewer's comments regrading a consistent use of ERA-5 throughout the paper, and therefore we have adjusted the methods section accordingly. In the revised version, the stratospheric extreme events have been detected using the ERA-5 data (which allows a direct comparison with the dates detected by ERA-Interim). Using ERA-5 for events detection does not change the results of this paper.

3) Next, since the ERA5 is used to initialize the ECMWF reforecasts, and since the two share aspects of their modeling components, independence in the comparisons is hard to justify. Again, this aspect limits the applicability of the results of this work to other forecast systems and reanalyses.

Although the ECMWF reforecasts are initialized from ERA5 data, they evolve from the reanalysis with time. Therefore, independence in the comparison is not needed for the verification of the forecasts. This study focuses on the ECMWF model from the reasons described in the previous answers, which allows an analysis of the model bias in the 4-weeks following the forecast initialization.

4) Next, the authors also comment frequently on the limited sample size from ERA5 for their results. This facet factors into their significance testing and other conclusions (e.g., Fig. 5). If sample size is too small, why should we trust the results? I am not saying that the limited sample size is a game-ender for the paper (trust me - this is a constant issue with my own work!). But, to use this concern over and over again in the manuscript as a caveat raises questions as to whether or not the findings are just an artifact of a short sample size.

Furthermore, we emphasize the variability among the events and its implications on event predictability. Focusing on the inter-variability of these events, given the small sample size limitation, complements the results shown by the composites of zonal wind and cyclone frequency (e.g., Figures 2 and 3) and provides a more detailed perspective on the impact of these events - without making any preliminary assumptions on their systematic bias. In this way, we are able to take a more careful path and overcome the limitation of the historical record. **We thank the reviewer for pointing out this important topic. Indeed, studies of atmospheric variability, and stratospheric variability in particular, are limited by the small sample size of the observational record. One possible way to increase the sample size is using atmospheric models, that are run for longer time periods compared to the historical record, and thus able to produce a larger number of events (e.g., Afargan-Gerstman et al., 2022). The extent to which we can assess biases is limited by the sample size (e.g., Lawrence et al., 2022; Domeisen et al., 2020). Despite a relatively small sample size (as in this study), these studies are able to assess model biases in the stratosphere across a wide range of subseasonal forecast systems when the results do show evidence of a systematic bias. To overcome the issue of the small sample size in Section 3.7, we take a different approach and combine all events with a canonical surface response (based on a comparison with the reanalysis) and analyze the individual ensemble members (in total 140 members; see Figure 10).**

Furthermore, we emphasize the variability among the events and its implications on event predictability. Focusing on the inter-variability of these events, given the small sample size limitation, complements the results shown by the composites of zonal wind and cyclone frequency (e.g., Figures 2 and 3) and provides a more detailed perspective on the impact of these events - without making any preliminary assumptions on their systematic bias. In this way, we are able to take a more careful path and overcome the limitation of the historical record.

References:

- **Afargan-Gerstman, H., Jiménez-Esteve, B., and Domeisen, D. I. (2022). On the Relative Importance of Stratospheric and Tropospheric Drivers for the North Atlantic Jet Response to Sudden Stratospheric Warming Events. Journal of Climate, 35(19), 2851-2865.**

- **Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., ... and Taguchi, M. (2020). The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere-troposphere coupling. Journal of Geophysical Research: Atmospheres, 125(2),**

e2019JD030923. https://doi.org/10.1029/2019JD030923

- **Lawrence, Z. D., Abalos, M., Ayarzagüena, B., Barriopedro, D., Butler, A. H., Calvo, N., de la Cámara, A., Charlton-Perez, A., Domeisen, D. I. V., Dunn-Sigouin, E., García-Serrano, J., Garfinkel, C. I., Hindley, N. P., Jia, L., Jucker, M., Karpechko, A. Y., Kim, H., Lang, A. L., Lee, S. H., Lin, P., Osman, M., Palmeiro, F. M., Perlwitz, J., Polichtchouk, I., Richter, J. H., Schwartz, C., Son, S.-W., Statnaia, I., Taguchi, M., Tyrrell, N. L., Wright, C. J., and Wu, R. W.-Y.: Quantifying stratospheric biases and identifying their potential sources in subseasonal forecast systems, Weather Clim. Dynam., 3, 977–1001, 2022.** https://doi.org/10.5194/wcd-3-977-2022

5) Finally, I was disappointed that the paper did not investigate any physical reasoning for why the storm tracks change as they do in reanalysis vs the reforecasts. The authors mention a few times that their results are "consistent with" previous studies, which is good. But, the reforecasts and their multiple ensemble members offer a fantastic opportunity for the authors to address the "why." They could explore changes in wave fluxes, baroclinicity, jet stream dynamics, etc. and provide an idea of why the stratosphere is influencing the storm tracks the way it is. I think this is a missed opportunity with this paper, thus making it contribution less novel than it otherwise could be.

We thank the reviewer for these suggestions. We have made changes to the manuscript to address the physical aspect of the storm track change in the reforecasts, compared to the reanalysis. Specifically, we have added a new subsection (3.7) to analyze and discuss the dynamical aspects of successful and unsuccessful predictions after SSW and strong vortex events. For this purpose, we explore the changes in troposphere, represented by the mean sea level pressure (MSLP) anomaly, and in the lower stratosphere, represented by the geopotential height anomaly at 100 hPa (Z'100).
Predictability of the downward impact after extreme stratospheric events strongly differs among events, even of the same type (e.g., Domeisen et al., 2020; Wu et al., 2022). The reasons for the observed differences in the predictability are not yet resolved, and often require an analysis from a case-by-case perspective, as was done for example for the 2018 SSW events (e.g., Karpechko et al., 2018, Kautz et al., 2020).
To determine the source of predictability of the downward response, we analyze the physical difference between SSWs that were "successfully"/"unsuccessfully" predicted (based on a criterion for a successful prediction, defined as forecasts in which the majority of ensemble members predict the observed sign of response in the midlatitude North Atlantic box). To guarantee a consistent surface response, only events with a canonical downward response are analyzed. We find that ensemble members with

a successful prediction of the canonical downward influence after SSW event differ from unsuccessful members mostly in their representation of tropospheric circulation anomalies after SSW events: the unsuccessful members do not predict the North-South dipole pattern in the Atlantic that corresponds to a negative NAO pattern (as shown by the MSLP patterns in Figure 10), despite well capturing the circulation anomalies in the lower stratosphere (as shown by the Z'100 patterns in Figure 10). These results indicate that the troposphere plays a dominant role in the downward impact of stratospheric anomalies after SSW events. Following strong polar vortex events, however, members with successful predictions differ from unsuccessful members in both their tropospheric and lower stratospheric anomalies.
Overall, this analysis sheds light on our conclusions presented in the manuscript, regarding the role of the tropospheric circulation in determining the predictability of the downward response of extreme stratospheric events. We have added this analysis to the manuscript (Section 3.7; Figure 10).

References:

- **Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., ... and Taguchi, M. (2020). The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere-troposphere coupling. Journal of Geophysical Research: Atmospheres, 125(2), e2019JD030923.** https://doi.org/10.1029/2019JD030923
- **Karpechko, A. Y., Charlton-Perez, A., Balmaseda, M., Tyrrell, N., and Vitart, F. (2018). Predicting sudden stratospheric warming 2018 and its climate impacts with a multimodel ensemble. Geophysical Research Letters, 45(24), 13-538.** https://doi.org/10.1029/2018GL081091.
- **Kautz, L. A., Polichtchouk, I., Birner, T., Garny, H., and Pinto, J. G. (2020). Enhanced extended-range predictability of the 2018 late-winter Eurasian cold spell due to the stratosphere. Quarterly Journal of the Royal Meteorological Society, 146(727), 1040-1055. Wu, R. W. Y., Wu, Z., and Domeisen, D. I. (2022). Differences in the sub-seasonal predictability of extreme stratospheric events. Weather and Climate Dynamics, 3(3), 755-776.** https://doi.org/10.5194/wcd-3-755-2022

Other Comments

1. The acronym "SPV." The use of this acronym is confusing - it is normally used to mean "stratospheric polar vortex" in many other papers. Furthermore, I don't find that the acronym is necessary in the work - "strong vortex events" is clear enough and not overly long. I recommend that the authors reconsider using this acronym.

The use of the acronym "SPV" for "strong polar vortex" can be found in the literature (e.g., Oehrlein et al., 2020, Díaz-Durán et al., 2017). However, to avoid confusion due to the different uses of this acronym, we corrected "SPV" to "strong vortex event" throughout the manuscript. In specific places we use the acronym "SV" for "strong vortex".

- **Oehrlein, J., Chiodo, G., and Polvani, L. M. (2020). The effect of interactive ozone chemistry on weak and strong stratospheric polar vortex events. Atmospheric Chemistry and Physics, 20(17), 10531-10544.**
- **Díaz-Durán, A., Serrano, E., Ayarzagüena, B., Abalos, M., and de la Cámara, A. (2017). Intra-seasonal variability of extreme boreal stratospheric polar vortex events and their precursors. Climate Dynamics, 49, 3473-3491.**

2. Lines 154-155. I don't understand this sentence. How is the "response in ERA5"... "stronger in ERA5?"
We rephrased this paragraph and removed this sentence.

3. Line 221. Either the results are statistically significant or they are not - they cannot be "partly significant."
Thank you for pointing this out. We have rephrased that sentence to clarify that the results in that case (difference in cyclone intensity) are not significant in ERA5 (lines 222-223).

4. Lines 281-283. Is this a "result" or "finding" that is unique to this work? I think that finding has already been shown in many past works and is also based on the fundamentals of what the jet stream is.
We rephrased this paragraph to emphasize that our analysis is consistent with previous studies, in context of the expected stratospheric impact. However, we emphasize a possible overconfidence of the model with respect to reanalysis to predict the canonical response after SSW events - a topic which has received less attention in the literature.

5. Lines 291-293. How would the authors propose to increase the sample size to meet their objective of determining the robustness of the results? (See my comments above as well.)
One possibility for increasing the sample size is to use a modeling study, hence to run a model for a longer time and generate more SSW events (e.g., Afargan-Gerstman et al., 2022). However, such procedure is less relevant when analyzing the ECMWF forecasts, as done in the current study. To overcome this issue, we take a different

approach in Section 3.7, and perform our analysis on the individual ensemble members (140 members for each event type) rather than an event-based analysis (14 events for each event type). By taking this approach, we are able to analyze larger sample size for each event type, and to gain more statistical insights on their predictability and hence dynamical aspects.

6. Figures 2 and 3. How is significance tested exactly for the reforecasts? What is the null hypothesis?
In Figures 2 and 3 significance is tested based on a Student's t-test. The null hypothesis (H0) is there is no difference between the means of these two variables (i.e., zonal wind anomalies and 0). Significance is tested for each grid point. An additional and more detailed significance testing is performed in section 3.4, where we investigate how the average cyclone life cycle characteristics depend on the extreme states of the stratospheric polar vortex. In Figure 5 and Figure 6, the confidence interval is obtained from a bootstrapped distribution of median latitudes (based on 1000 random resamples of the tracks with replacement).

7. Figure 7. "Successfully" is spelled incorrectly in the y-axis labels of panels (b) and (d). Also, it is unclear what an "increase of cyclone frequency anomaly" means. Is it that it is a positive anomaly, or that the anomaly actually gets more positive over some time?
We have corrected the typo. As for the meaning of "increase of cyclone frequency anomaly", this term refers to a positive anomaly in cyclone frequency. Red bars in Fig. 7a indicate the proportion of ensemble members that show an average increase in cyclone frequency over the selected region, whereas blue bars indicate a decrease.

8. Code and data availability. The authors have not provided a public-accessible repository where their code is available. Please set up a Github and place your code on there for transparency and accessibility.
We thank the reviewer for this comment. We have created a public-accessible Github repository for the code and datasets (`https://github.com/hillaag/downward_impact_analysis_tools_for_S2S.git`). We now specify this information under the Code Availability section.