Response to the reviewers of wcd-2021-84: Differences in the Sub-seasonal Predictability of Extreme Stratospheric Events

Dear Co-Editor Yang Zhang,

On behalf of all authors I would like to submit the revised version of the original article "Differences in the Sub-seasonal Predictability of Extreme Stratospheric Events" including an annotated manuscript and a modified version of the manuscript. During the revision, a number of changes have been made to the manuscript to satisfy the requests of the reviewers. Please find the summary of the changes and the detailed responses to the reviewers below.

Best regards,
Rachel Wu

We would like to thank both reviewers for their helpful comments and suggestions for our study. These have been included into the manuscript (see changes indicated in **bold** in the annotated manuscript). Please find below the detailed responses (in blue) to the reviewers' comments and suggestions. All line indications refer to the new (annotated) version of the manuscript unless specified. The main changes to the manuscript are listed here:

Changes in Sections:

- Section 2.2 *Skill measures* is now Section 2.3 and Section 2.3 *Definition of stratospheric events* is now Section 2.2

Changes in Figures:

- Figure 1: Separated into two panels, one for deceleration events, one for acceleration events
- Figure 3: Grouped every 2 LTGs together, reduced from 6 panels to 3 panels
- Figure 3, 6, 8, 9: now no longer distinguish weak and strong magnitude events
- Figure 4: Now showing 3 panels, one panel for each metric
- Figures 5 and 7: $\overline{u}_{yy}$ was in wrong units in the original manuscript, now corrected with the correct units
- Figure 6, 8, 9: Updated with $\overline{u}_{yy}$ averaged over day -10 to -1 instead of using $\overline{u}_{yy}$ at day 0
- Figure 7: we combined every two LTGs rather than showing all LTGs in the original manuscript

Main topics of reviewer questions:
We here list a few more general answers to major points that came up during the review.

1 **Chosen threshold for event definition vs. sample size:** We have carefully evaluated the chosen threshold for the defined acceleration and deceleration events with respect to the balance between the 'extremeness' of the events and the sample size. A more extreme threshold (i.e. a higher percentile) yields fewer and more extreme events, while it strongly reduces the sample size. Our goal is to have a larger sample size than what we obtain for limiting the sample to SSW and strong vortex events, but to keep the threshold sufficiently high to detect only events that can still be considered strong or even extreme, hence choosing the event threshold is a delicate balance. Several studies have been performed on the limited sample of SSW and strong vortex events, and given the limited amount of observational data it is difficult to understand the mechanisms for these events beyond case studies and simplified models. We therefore expand the sample to deceleration and acceleration events in order to better be able to find

commonalities between events, while still allowing for these events to be strong enough to exhibit many of the same mechanisms and characteristics that are observed for SSW and strong vortex events. It has to be noted that some of the strongest SSW events are possibly driven by mechanisms – such as resonance – that are highly nonlinear, and which may not be fully reproduced in the prediction system used here. The goal of our study is not to investigate the mechanisms for these very extreme events, but the basic ingredients such as wave driving and the background state that remain important ingredients worth investigating for all cases, and our study therefore aims to shed light on the nature of these events by balancing the sample size and the extreme nature of these events. We have elaborated further on this point in the detailed answers below.

2 **Nonlinear nature of stratospheric events:** As noted in point 1 above, the most extreme SSW events may exhibit different and/or additional mechanisms that are highly nonlinear. For example, some of the most extreme SSW events have been suggested to be driven by resonance. However, all deceleration events including strong SSWs require a minimum threshold of wave flux input into the stratosphere, and furthermore the stratospheric background state is often found to be critical for wave amplification. Therefore, these two ingredients are two clear starting points for our study, i.e. to investigate if the model is able to reproduce the wave flux and the stratospheric state in order to be able to reproduce stratospheric events. By no means do we intend to suggest that stratospheric events are driven by purely linear mechanisms, but it is rather the deviation from linear relationships that we are interested in. Nevertheless, we find relationships that are close to linear in terms of their predictability (rather than their dynamics), e.g. the relationship between the CRPS and the wind change (Figure 3). We have elaborated further on this point in the detailed answers below.

3 $\overline{u}_{yy}$ **as a measure of the background state:** We demonstrate with scatter plots in Figures 6-10 in this document that $\overline{u}_{yy}$ correlates well with $\overline{u}_{zz}$ (the third term of $\overline{q}_y$), and that $\overline{u}_{yy}$ correlates well with $\overline{q}_y$. In Lines 183-185 in the manuscript, " Other than being a reasonable indicator for the refractive index, $\overline{u}_{yy}$ is a measure of the sharpness of the edge of the stratospheric polar vortex, thus also a measure of the strength of the initial vortex state." Thus, in the light of the scatter plots and correlation we include in this response, we decided to keep our interpretation of $\overline{u}_{yy}$ as a measure of the background state of the vortex, and a reasonable proxy for the refractive index, for the region we consider. To filter out high frequency variations in $\overline{u}_{yy}$, we have modified our $\overline{u}_{yy}$ index by using a 10-day averaged value of $\overline{u}_{yy}$ at day -10 to -1 instead of using the value of $\overline{u}_{yy}$ at day 0 that was originally used in the study.

Reviewer 1

General Comments

The manuscript addresses an important gap in the subseasonal-to-seasonal (S2S) community - an investigation of how predictable rapid acceleration and deceleration polar vortex events are in the ECMWF subseasonal forecasting system. Quantifying this predictability is important, as changes in the strength of the Northern Hemisphere stratospheric polar vortex typically precede changes in winter weather regimes in the troposphere. The authors find that, while the ECMWF performs well in terms of the driving mechanisms for these acceleration/deceleration events, it cannot capture the magnitude of the most extreme events, a finding common to other prediction systems. This discrepancy in magnitude is especially true for the wave fluxes, which are underestimated in the model. Altogether, the analysis of the model and comparisons with reanalysis is done generally well, and the authors have identified a couple of key metrics that could be assessed for these events. These two metrics - meridional heat flux and a proxy for the index of refraction - could be useful in future assessments of subseasonal forecasting systems and their stratosphere-troposphere coupling mechanisms. The statistics shown are valid and comprehensive, though admittedly numerous and could be streamlined. I think that the conclusions follow the analyses conducted, though a bit more on the mechanistic framework and some more spatial-dependent analyses could help the paper. As such, I am suggesting that the work undergo major revisions before acceptance.

$\overline{v'T'}$ \ $\overline{u}_{yy}$	-30 to -21	-20 to -11	-10 to -1	0 to 9
0 to 9	0.47/0.00	0.44/0.00	0.30/0.02	-

$\overline{v'T'}$ \ $\overline{u}_{yy}$	-30 to -21	-20 to -11	-10 to -1	0 to 9
0 to 9	-	-	-	-

Figure 1: Correlation of 10-day integrated sum of $\overline{v'T'}$ over day 0 to 9 and the 10-day averaged $\overline{u}_{yy}$ index at different lags, e.g. -30 to -21 denotes 30 to 21 days before the start of the events. The Pearson coefficient (first value) and p-value (second value) are only shown for significant correlations. The top table shows the values for deceleration events and the bottom table for acceleration events.

Thank you for your insightful evaluation of our manuscript. We are responding to your detailed comments below.

Specific Comments

- **Interdependence of Refractive Index and Wave Forcing.** The authors examine mechanisms and drivers that could explain strong acceleration and deceleration events. To do this, they have examined the index of refraction and meridional heat flux. However, the authors indirectly treat these two metrics as independent and look at their evolution separately. In fact, the authors treat the index of refraction as a measure of the "background state of the stratosphere" (Line 245). However, these two variables are a function of each other. While initially the refractive index may facilitate wave propagation, the breaking of waves in the stratosphere and the changes in the zonal winds and heating profiles caused by these breaking waves will alter the refractive index, which in turn influences future wave breaks. So, it is hard to keep the two metrics completely separate. Have the authors considered this interdependence and thought of ways to address it? For example, if a model poorly handles wave fluxes 25-30 days before an observed event, can we actually use the simulated index of refraction to assess its prediction of an event?

  Thank you for your comment. We agree that the two selected indices, i.e. the refractive index and the meridional heat flux, are related to each other. To address the problem of interdependence, we revisited the definition of the metrics and considered the averaging window at different time lags for the metrics, to better separate the indices.

  We compare the correlations of 10-day averaged $\overline{u}_{yy}$ index at different lags with the 10-day integrated sum of $\overline{v'T'}$ over day 0 to 9 for both deceleration and acceleration events at different lags in Figure 1. In the top table, for deceleration events, we find significant positive correlations in $\overline{u}_{yy}$ with the integrated $\overline{v'T'}$ over day 0 to 9 whenever $\overline{u}_{yy}$ leads $\overline{v'T'}$, i.e. for $\overline{u}_{yy}$ averaged over day -30 to -21, day -20 to -11 and day -10 to -1. In the bottom table, for acceleration events, no significant correlation is found between $\overline{u}_{yy}$ and the integrated $\overline{v'T'}$ over day 0 to 9.

  As such, we agree that we should not treat the two indices as independent for deceleration events. To minimise the dependence of the two indices, we choose the time lag that exhibits the lowest correlation between the two metrics, i.e. day -10 to -1 for $\overline{u}_{yy}$, as the new averaging window for the $\overline{u}_{yy}$ metric. We have also added Lines 189-194 in the manuscript to clarify that these indices should be treated with care and that the two metrics are not fully independent.

- **Spatial Analyses.** The manuscript studies all events and their forcings in a zonal-mean framework. That approach is a classical way to look at stratosphere-troposphere coupling, but emerging evidence points to the importance of polar vortex morphology and tropospheric source regions of waves for understanding circulation anomalies in the troposphere and stratosphere. As such, spatial distributions of meridional heat flux (at a given isobaric level or even as a cross-section) could be very informative to understand whether the models initiate the waves in the right places. For example, climatologically, vertical wave propagation has two major hotspots during boreal winter: (a) Siberia and (b) Scandinavia

/ Northern Europe. However, other forecasting systems possess biases on where these hotspots are because of their representation of planetary-scale waves. How does the ECMWF perform in this context, and specifically during strong acceleration or deceleration events? Is one region better represented than the other? Also, what about the morphology of the stratospheric polar vortex? How is that different in the lead up to strong and weak acceleration events, and could that be a predictive element? I am offering two suggestions here, but others are possible. My main point is that I would like to see more multi-dimensional analyses in addition to the zonal-mean metrics (which are important!).

Thank you for your comments and suggestion. We agree that regional analysis of wave activity would be interesting and a nice addition to our results. As an extra analysis, we have separated the reanalysis data of meridional heat flux into contributions from four regions that are selected based on the existing literature and on our own analysis. Specifically, based on the meridional heat flux composite of deceleration events (Fig. 2a), we divided the 45-75°N latitude region equally into four regions as indicated in Figure 2: (a) Northern Europe (40°W - 50°E), (b) Siberia (50°E - 140°E), (c) North Pacific (140°E - 130°W) and (d) North America/ Greenland (130°W - 40°W). Comparing the composite of deceleration events to that of the Nov to Mar average, we see positive heat flux anomalies in three regions, namely, Northern Europe, Siberia and the North Pacific, and negative heat flux anomalies in the region North America/ Greenland during deceleration events. The composite for acceleration events also shows similar patterns. Thus, we choose to average the heat flux over the same four regions for both deceleration and acceleration events to examine the predictability of the wave activity captured by the model at different lead times.

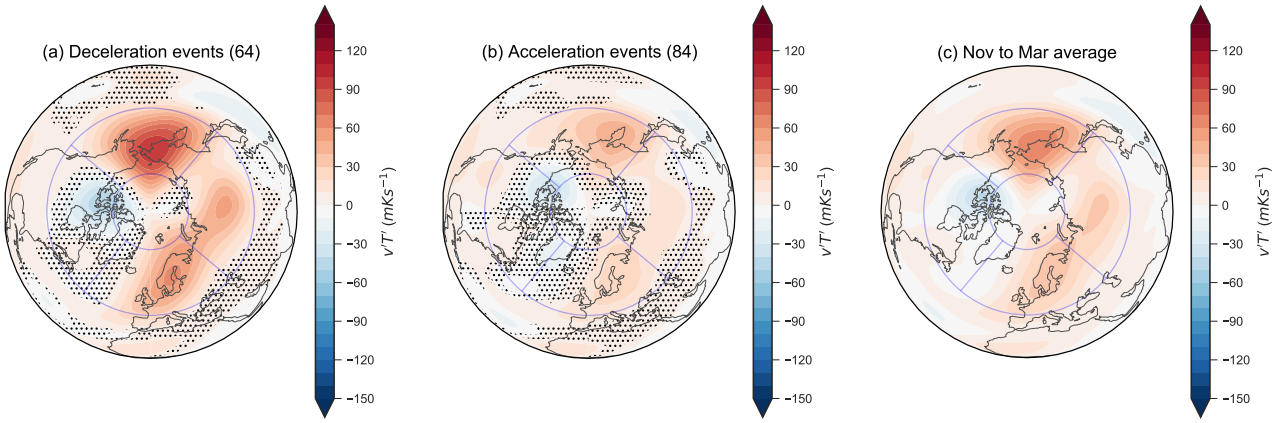

Figure 2: Composite $v'T'$ at 100 hPa averaged over day 0 to 9 (during the event window) of (a) Deceleration events, (b) Acceleration events and (c) Nov to Mar average in reanalysis. Blue lines mark the regions of investigation. Northern Europe (40°W - 50°E), Siberia (50°E - 140°E), North Pacific (140°E - 130°W) and North America / Greenland (130°W - 40°W). Numbers in bracket indicate number of events in the composite and unhatched regions in (a) and (b) indicate statistically significant different from (c) by a t-test.

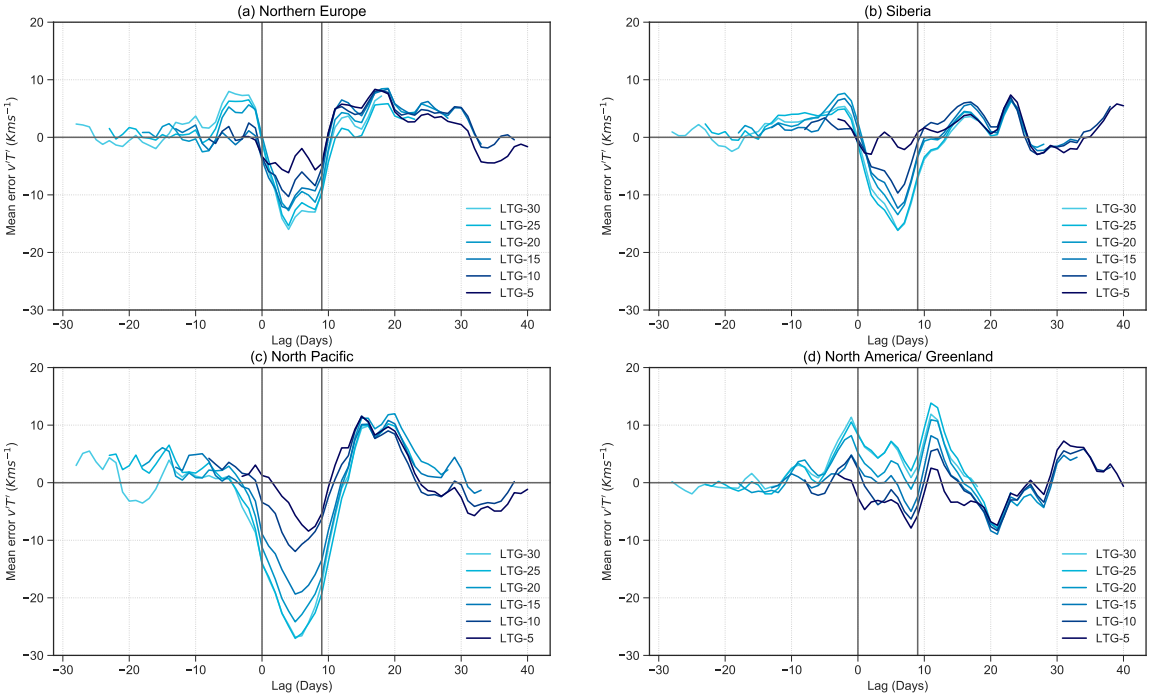

Figure 3: Mean error of the composite $v'T'$ at 100 hPa for deceleration events over (a) Northern Europe, (b) Siberia, (c) North Pacific and (d) North America / Greenland predicted by the hindcasts at different lead times.

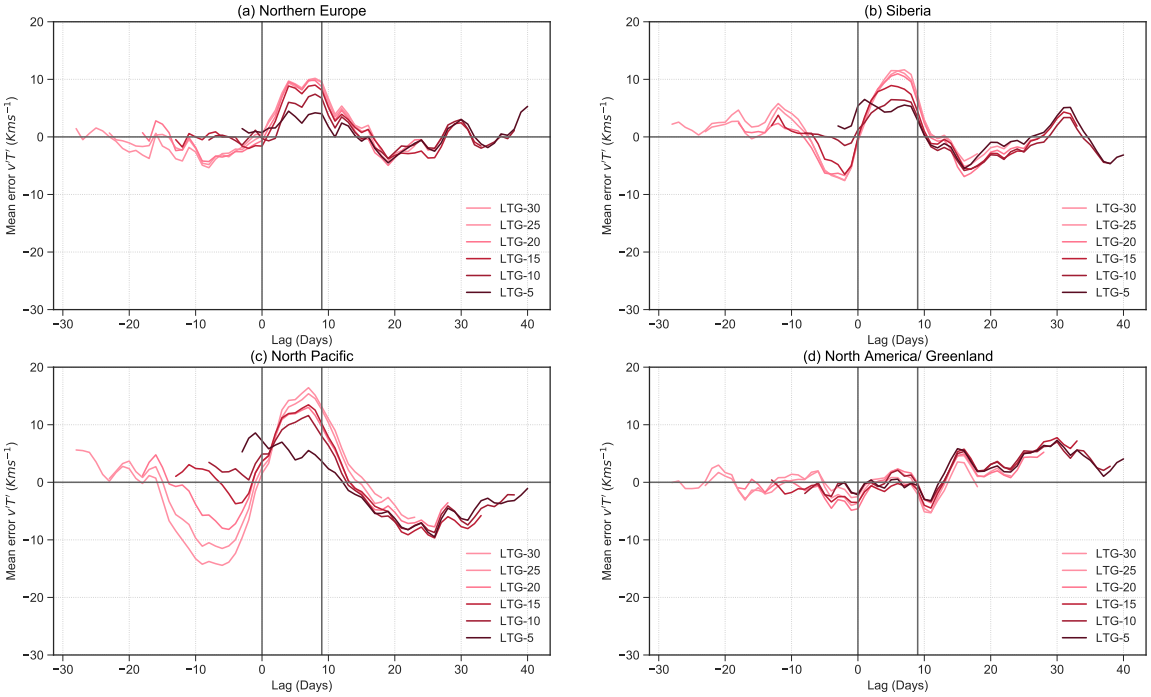

Figure 4: Same as Fig. 3 but for acceleration events.

Fig. 3 shows the mean error of the averaged $v'T'$ for the four regions for the deceleration events in the hindcast with respect to the reanalysis. The mean error is largest in the North Pacific, while it is of similar magnitude for Northern Europe and Siberia, with larger errors for longer lead times. The model generally underestimates $v'T'$ during the event window in these three regions. The error is comparably smallest for the North America/ Greenland region. Since the averaged heat flux is negative in the region, the positive mean errors at long lead time indicate that the averaged $v'T'$ in the model of the region is not as negative as that in reanalysis. At short lead time, the mean errors for North America/ Greenland become negative, indicating that the $v'T'$ for the region in the model becomes more negative than in reanalysis.

The same analysis is done for the acceleration events (Fig. 4). The mean error in the North America/ Greenland region is close to zero for all LTGs, while the model overestimates the $v'T'$ for the North Pacific, Northern Europe and Siberia. The error in the North Pacific is larger than for Northern Europe and Siberia, as observed for deceleration events. However, the errors in the North Pacific for acceleration events are more comparable in magnitude to Northern Europe and Siberia than for deceleration events.

To sum up, $v'T'$ in the North Pacific contributes more than the other regions in terms of the errors in $v'T'$ and the North America/ Greenland region contributes less. However, all regions contribute to the $v'T'$ errors for the deceleration events and all regions except the North America/ Greenland regions for the acceleration events. A more detailed analysis on the regional origin of errors will have to be performed in the future.

To describe the new results in the manuscript, we have added Lines 406-410: 'We further investigate the regional origin of the $\overline{v'T'}$ errors by dividing up the regions of heat flux origin into Northern Europe, Siberia, North Pacific and North America/ Greenland (Fig. A4). All regions contribute to the errors, with the largest contribution coming from the North Pacific, and smaller contributions from Northern Europe and Siberia (Fig. A5 and A6). Additional analysis is needed to further understand the origin of the $\overline{v'T'}$ errors.' We have also added the figures into the Appendix of the manuscript as Fig. A4 to A6 and the description to Fig. A4 in Lines 467-476.

- **More Justification for Choice of Events.** The authors provide definitions for their strong and weak magnitude events as being above their respective 60th percentiles. However, I am unsure why this percentile is chosen other than that threshold is used in other works. In fact, I do not consider the 60th percentile as "extreme" as the title of the manuscript indicates. I would like the authors to provide more details on the choice of this threshold and also how sensitive their analyses and conclusions would be if the value was shifted to the 75th or 80th percentiles.

  [Please see also major point 1 at the beginning of the reviewer response] Thank you for your comment. The chosen threshold is based on obtaining strong events while also having a large enough sample that is larger than the limited sample of SSW or strong vortex events. We have made an effort to put our chosen threshold in context by comparing with previous studies that define similar events. For example, the threshold that we use for strong magnitude events has a similar magnitude as the definition used in Birner and Albers (2017), who classified *sudden stratospheric deceleration events*. We have, therefore, re-framed our event definition in order to better justify our choice of threshold for events and to what extent the classified events can be considered extreme events.

  In Birner and Albers (2017), the threshold is chosen based on the standard deviation of the deseasonalised daily zonal mean zonal wind of the time period they considered, where the threshold for a 10-day event is defined by multiplying the standard deviation of the daily value by 10 days. The 10-day threshold they use is 20 ms^{-1} over 10 days, which corresponds to a daily zonal mean zonal wind change of around 2.2σ per day, which is equivalent to a daily wind change of around the 98th percentile. Following Birner and Albers (2017), we compute the standard deviation of the time series of the deseasonalised zonal mean zonal wind of our time period of consideration, which is around 1 ms^{-1}/day. Therefore, our definition of strong acceleration and deceleration events of 16.9 ms^{-1} over 10 days and 24.6 ms^{-1} over 10 days corresponds to daily wind changes between the 95th and 99th percentile values (1.69σ and 2.46σ) in NH Nov-Mar. Therefore, we keep our original definition of events. We have adapted Lines 103-125 of the manuscript to better explain our choice of threshold.

  Having revisited the figures in the manuscript, in most of the figures of the original manuscript, we

distinguish between weak and strong magnitude events. We found that this distinction is not needed for most of our conclusions, and we draw conclusions from most of the figures using the entire event magnitude spectrum, i.e. considering all events together rather than just the strong magnitude events. However, in Figure 4 and 7, the strong magnitude events need to be distinguished to convey the messages from the figures. As such, other than Figure 4 and 7, we have modified all the other figures to not indicate the difference between weak and strong magnitude events.

As we now no longer distinguish strong magnitude events in most of the figures, it will only affect Figure 4 and 7 if a higher percentile is used as the threshold of the events. As a sensitivity test, we repeated the same analyses but using the 80th percentile. The results still apply for Figure 4 and 7. However, when using the 80th percentile as the threshold, much fewer events are detected and the number of identified strong magnitude events become comparable to the number of SSW events. We have therefore kept the original percentile definition and elaborated on it in Lines 116-125 in the manuscript.

- **Complexity of Figure 7.** I understand the motivation of looking at multiple lead times and ensembles when studying these different events and comparing their features to reanalysis. However, Figure 7 has seven differently colored lines (six of which are different shades of blue), two different line styles, and six colors of shading per panel. I found it difficult to differentiate the different blue colored lines, especially since many of them overlap each other. I think the authors should consider simplifying these figures by, for example, reducing the quantity of lines. Since we already know that the models improve with shorter lead times from the other previous analyses, can the same message come across with just LTG-25, LTG-10, and LTG-5? Are all the shading colors needed? Again, I am thinking of ways of making this figure more accessible and cleaner without losing its meaning.

  Thank you for the comment. We have modified Figure 7 by plotting the average of every two LTGs to reduce the number of colours and lines to make the figure cleaner.

Technical Corrections

- **Lines 1-2.** The phrase "associated with an anomalously weak or strong polar vortex" is oddly placed. Please consider removing this phrase.

  Thanks, phrase removed.

- **Lines 10.** Please add a semicolon after "behaviour."

  Thanks, change made.

- **Lines 10-11.** The wording following "that is" reads awkwardly. Please consider revising.

  Thank you for the comment. The sentence is now reformulated in Lines 8-10.

- **Lines 34-35.** How does the strong latitudinal temperature gradient drive radiative cooling in the stratosphere? Isn't the radiative cooling a function of the (lack of) solar insolation during winter months?

  Thank you for this comment, the sentence was misleading. This sentence has been modified and combined with the following sentence and now reads "On the other hand, when wave activity is weak and the SPV is relatively undisturbed, the vortex strengthens on radiative timescales (Limpasuvan et al., 2005; Hitchcock and Shepherd, 2013)." in Lines 33-34.

- **Lines 39.** Please add "Major" before "SSW events."

  Thanks, change made.

- **Figure 1.** I suggest that the authors break this figure into two panels: one for the deceleration/SSW events and the other for the acceleration/strong vortex events. As presented, the one plot has a lot of information and is too cluttered to understand fully. Moreover, is **Line 197** correct? When I examine the figure, I see the blue line (median for deceleration events) higher than red line, indicating a higher magnitude error for deceleration events, not the other way around. Maybe it is just hard to see in the figure (for me), but could the authors check this and perhaps explicitly state the values of the medians just to make sure?

Thank you for the comment. We have now split Figure 1 into two panels as suggested. Thank you for spotting the error at Line 197 in the original manuscript. The median for deceleration events is larger than acceleration events in LTG-5, which indicates that mean error for deceleration events is higher than acceleration events for all lead times. We have now replaced the sentence with '..., we also find that deceleration events are associated with larger errors than acceleration events at all lead times.' in Line 218-219.

- **Line 203.** Please add "wind changes" after "magnitude" to make clear what the magnitude represents.

  Thanks, change made (Line 224).

- **Figure 2.** In the caption, please change "brackets" to "parentheses."

  Thanks, change made.

- **Lines 213-216.** I read this sentence several times, and I still do not understand what it is saying about the gray diagonal line. Please consider rewriting.

  Thank you for pointing this out. We have added the following sentence on Lines 156-159 in Section 2.3: "As the CRPS is given by the difference between the predicted and observed distribution, if all ensemble members in a hindcast predict an event magnitude of 0 ms^{-1}, i.e. close to a climatological state where the wind stays relatively constant during a 10-day window, the CRPS of this hindcast will be equal to the observed event magnitude itself.". We here aim to explain how the CRPS of a hindcast that predicts a climatological state will be equal to the event magnitude of the actual event. Then, in Lines 233-235 of the Results section, we shortened our description on the grey diagonal line and stated directly how the diagonal line is used as a reference to compare the skill of the data points to a climatological forecast.

- **Lines 228-229.** This line starting with "For instance" is a fragment and should be corrected.

  Thank you for catching this. We have modified the line to "Some events, for instance, the two extreme SSW events with magnitudes of over 60 ms^{-1} (marked by yellow stars in Fig. 3), retain a large CRPS and deviate from the linear fit in the direction of the diagonal line." in Lines 245-247.

Reviewer 2

General Comments

This paper examines predictability of wind deceleration and acceleration events using the ensemble hindcasts of the ECMWF for the period of 1998-2018. The variability and predictability of those events are examined according to the magnitude change of the zonal-mean zonal winds, its meridional curvature at 60N, 10 hPa and eddy heat fluxes in the lower stratosphere. It is found that the model can reasonably predict the acceleration events but unable to reproduce extremely deceleration events, which effectively the SSWs. The inability of the model to produce SSWs is linked to weaker-than-observed eddy heat fluxes in the lower stratosphere within the same 10-day interval.

The evaluation of the statistical representation of acceleration and deceleration events are interesting, e.g. the model continues to underestimate the long tails associated with deceleration events, even at short lead times; how the distributions of various quantities compared with reanalysis data sets. I however have major concern in terms of the dynamical reasoning. See comments below for details.

Thank you for your comments and insightful evaluation of our manuscript. Below are point-by-point responses, and we also added further discussion in the revised manuscript in response to your comments.

Major Comments

- The mechanisms that the authors identified are entirely consistent with the linear theory, which is adequate in explaining the climatological behaviour of stratosphere wave mean-flow interaction and

polar vortex variability, but not sufficient in explaining the SSWs. Thus, the title of the paper does not match its content or key results.

[Please see also major point 2 at the beginning of the reviewer response] Thank you for your comment. We fully agree that linear theory is not sufficient to explain SSW events and that processes that are described by nonlinear theory, such as resonance, would be needed to – in particular – fully explain very strong SSW events. We do not aim to, in our paper, to explain SSWs with linear theory. Instead, our aim is to trace the sources of the predictability of stratospheric extreme wind events in the model. As a first approximation, we investigate deviations from linear relationships between the predictability of events and their magnitude (e.g. Figures 3 and 9) and between precursors and event magnitude (Figure 6). The finding is that the stratospheric events follow linear relationships to some degree, but that especially extreme events – unsurprisingly given their nonlinear dynamics – deviate from the linear relationship.

The model in general captures the linear part of the wave activity forcing well for most events except for the strong magnitude events, implying that the model shows an inability to capture very strong wave activity, which might be related to the inability of the model to capture nonlinear processes. As such, (please also refer to the response to the next comment) we think our results are important for extreme stratospheric events, since the strong magnitude events we identified have magnitudes that are comparable to extreme stratospheric events. For more justification of the 'extremeness' of the events please see the answer to reviewer 1, point 3 ('More Justification for Choice of Events'). We have therefore decided to keep the title of our paper.

- The results presented shade little new insight onto the predictability of extreme stratospheric events, i.e. SSWs. This is mainly because the authors use upper and lower 60th percentiles of negative (or positive) deltaU within a 10-day window to define the deceleration (or acceleration) events, which is not the standard measure of extreme events. For instance, a normal distribution can approximately capture the 60 percentiles of generalized extreme value (GEV) distribution, but it would fail to model the long tails of the GEV, which normally corresponds to bottom or top 1-5 percentiles of a distribution. Thus, including small-magnitude events will result in better statistics but potentially hides the responsible mechanisms for the extreme events because the statistics provided by the 60 percentiles of a population is not representative of its extreme values.

  [Please see also major points 1 and 2 at the beginning of the reviewer response] Thank you for the comment. We would like to first clarify that by the term extreme stratospheric events, we are referring to both strong vortex events and SSW events (see Section 2.2 for more detailed definitions), which are commonly studied in predictability studies, as well as strong deceleration and acceleration events, which have a strong overlap with SSW and strong vortex events.

  We are aware that the definition of using the 60th percentiles of deltaU within a 10-day window might not be considered extreme. However, when we present our definition of extreme events in terms of wind change per day, our definition for deceleration and acceleration events correspond to wind change within the 95th and 99th percentiles, respectively, per day. (For more justification of the 'extremeness' of the events please see the answer to reviewer 1, point 3 ('More Justification for Choice of Events').) Our definition of the strong deceleration events is comparable to the definition of sudden stratospheric deceleration events or alternative definition of SSWs based on zonal-mean zonal wind tendency in (e.g. Birner and Albers, 2017; de la Cámara et al., 2019; Kim et al., 2017), where the threshold they use corresponds to around daily zonal mean zonal wind change of around 2.2 standard deviations per day and to a daily wind change of at around the 98th percentile.

  Following the definition of Birner and Albers (2017), we have computed the standard deviation of the time series of the deseasonalised zonal mean zonal wind, which is around 1 ms^{-1}/day. Our definition of strong acceleration and deceleration events, which is defined as events with wind change of 1.69 ms^{-1}/day and 2.46 ms^{-1}/day within a 10-day event window respectively, corresponds to 1.69 times of the wind changes of 95th and 99th percentile (1.69σ and 2.46σ) in NH Nov-Mar. We have added Lines 112-125 in the manuscript to better elaborate and justify our choice of event definitions.

  Furthermore, we agree that the strongest SSW events, although part of the sample presented here, are not fully captured in terms of predictability. In fact, we show that the strongest SSW events, e.g. the split events in 2009 and 2018, are very poorly captured in the prediction system (e.g. Fig. 3 in the

manuscript). We are not able to investigate these events in detail in this study due to their exceptional nature, and we have in this study rather focused on studying the general ingredients of predictability for wave acceleration and deceleration events. However, the extreme nonlinear events that are highly unpredictable are a highly interesting further topic to explore in a future study.

- The deceleration and acceleration events appear to include high frequency variability (i.e. < 5 days), the effect is readily seen in Figure A1. The authors need to either justify the extent to which the effects of these high-frequency waves on the polar vortex variability in relation to the SSWs or applying a lowpass filter to the 6-hourly data so that the variation within the 10-day window is truly relevant to extreme stratospheric events.

  Thank you for this comment. We agree it is undesirable to include high-frequency variability within the 10-day event window. We have actually taken this point into consideration when we identify the events. It was, however, not included in the original manuscript. Thank you for pointing this out.

  We have now included Lines 103-105, 'We also impose a criterion that the ratio of the maximum difference in between the maximum and minimum wind speed occuring during the 10-day event window has to be less than 1.2, to filter out high frequency variations.', to describe that a criterion is imposed when identifying events to avoid picking up events that include high-frequency variability. We have also included Harnik (2009) as a reference in Line 453 to comment on the effect of high-frequency variability on the polar vortex.

- The SSWs are known to involve nonlinear processes such as wave breaking, resonance, and internal wave reflection, some of which the model may have failed to capture. For instance, erosion and filamentation due to wave breaking can increase meridional curvature as well as enhance zonal winds at polar vortex edge via PV sharpening. Thus, wave forcing from below does not always result in a weaker polar vortex within a 10-day time window. As such, the meridional curvature term uyy is not a good measure of waveguide.

  [Please see also major point 2 on first page of reviewer response] Thank you for your comment. We fully agree that SSWs involve nonlinear processes. In our results, the model however shows limitations in producing extreme values of the eddy heat flux, which already indicates that the model does not fully capture processes such as resonance and internal wave reflection. This is, however, beyond the scope of our study.

  We also agree that wave forcing from below does not always result in a weaker polar vortex, and that the vortex can also be strengthened via PV sharpening. However, for longer timescales, i.e. the 10-day averages that are employed in our study, (e.g. Figure 5 and 6 in the manuscript), the 10-day integrated eddy heat flux during the 10-day event window has a strong positive correlation with the wind deceleration in the window, yielding an overall clear response. Hence, if we consider the integrated eddy heat flux in this time window (i.e. day 0 to 9), the wave forcing from below will most of the time result in a weaker polar vortex.

  In the light of your comment, we have added Lines 449-453 in the Discussion, 'For example, the ability of the model to capture the nonlinear dynamics, which are known to be relevant to SSWs with strong magnitude, has not been explored in this study. These nonlinear processes include the complex behavior of wave breaking, which depending on its exact location and temporal variability can have different effects on the polar vortex, for instance, high frequency wave activity can strengthen the polar vortex rather than weakening it (**?**). As such, ...'.

- A few multi-panel figures are too complicated and some of the panels are redundant. See specific comment below.

  Thank you for this comment, we have now simplified several of the figures in the revised manuscript following the comments by both reviewers. We have added a summary of the changes at the beginning of the reviewer response, please see above.

Specific Comments

- Line 4, page 1, delete "limit".

Thanks, change made.

- Line 10, page 1, “in a close to linear relationship”, it may not be appropriate to study extreme events using linear relationship.

  Thank you for the comment. In this case, we are describing the predictability of the events and comparing how closely the relationships between event magnitude and precursors, and between their predictability in the model resemble a linear relationship. By no means do we intend to imply that this relationship should be linear, or that the dynamics of the events should be linear. The regression line is used as a reference to compare the relationship between the wind change and the integrated eddy heat flux and the averaged $\overline{u}_{yy}$. We have now clarified this in Lines 196-199 in the manuscript.

- Line 13, page 1, “wave activity pulses”, I do not think that the authors studied wave activity pulses. The exact quantity studied is v'T' averaged within a 10-day window, which can contain only a part of wave pulse or multiple high-frequency wave pulses.

  Thank you for this comment. We have replaced "wave activity pulses" with "wave activity fluxes" in Lines 12 and 14.

- Lines 35 and 45, page 2, polar vortex can be strengthened via wave breaking and PV sharpening as well.

  Both sentences have been adapted, see Lines 33-34 and 44.

- Line 59-60, page 3, very good point Re initial stratospheric conditions, but the authors did not study this factor in the rest of the paper. Consider rephrase or remove the sentence.

  The second half of the sentence has been rephrased, in Lines 58-59, "..., suggesting that other factors, e.g., the background state of the stratosphere, might be important for successful predictions of SSWs.", as we investigate the predictability of the background state of the stratosphere in our study.

- Lines 123-134, page 5, using a fixed 10-day moving window to define the acceleration and deceleration events is problematic as it cannot properly differentiate high and low frequency variability thereby wave mean flow interaction. Harnik (2009) demonstrated that low frequency wave activity slows down the zonal winds while transient, high frequency wave pulses act to enhance the polar vortex.

  Thank you for this comment. One reason for using a 10-day event window, apart from having a consistent time window for defining wind changes, is to filter out high frequency signals (see Lines 103-111 in the manuscript). While we are aware that high frequency wave pulses also have an effect on enhancing the polar vortex, high frequency wave pulses are not the focus of this paper. To account for this finding, we have included a sentence about the effect of high frequency pulses into the manuscript in Line 453, citing the reference by Harnik (2009). In addition, we have evaluated a wide range of window widths as a sensitivity test. Figure A1, panel (a) shows that there is no major difference between acceleration and deceleration events in terms of their duration.

- Lines 141 -155, it is better to condense those roles/conditions and put them into Table 1. Also, the sample size for each subgroup in reanalysis are too small to establish robust statistics or to understand the relevant mechanisms. For instance, a strengthening of a polar vortex can be due to reduced upward wave forcing, PV sharpening via wave breaking, and/or enhanced meridional temperature gradient. It is nearly impossible to differentiate these causes merely based on 25 events.

  Thank you for your suggestion. We agree that combining those definitions into Table 1 will make it clearer for the reader. We have added the definitions as an extra column to Table 1. We have, however, kept the text that explains the criteria in the manuscript as we think it might allow readers to better follow the manuscript.

  About the sample size, we agree that we should not identify or differentiate causes and the mechanisms based on the small sample size from the reanalysis data. The goal of this study is to investigate and compare the predictability of stratospheric wind changes, and to identify possible model biases. We extend the definition from just SSW and strong vortex events, which have often been studied in terms of their predictability despite small sample sizes, to acceleration and deceleration events, which have a larger sample size. A further goal of the study is to give hints at the potential mechanisms that may not be sufficiently represented in the prediction model in order to give an accurate prediction. We therefore

evaluate whether the model captures the mechanisms as suggested from the literature and whether the model has biases in predictors that are known from the literature. The investigation, and thereby the sample size, is a balance between obtaining a sufficient number of events (that is larger than the observed record for SSW and strong vortex events) and the threshold for defining these events. We found that the chosen threshold gives us a number of events that allows for an investigation of the mechanisms, while still having events that are strong enough to compare to e.g. SSW events. [please see also major point 1 at the beginning of the reviewer response]

- Line 153, "the chosen threshold . . . ", at which pressure level and latitude?

  Thank you for pointing this out. We have replaced "the chosen threshold . . . " with the information on pressure level and latitude as "the chosen threshold value at 60°N and 10 hPa is 41.2 m/s" in Line 138.

- Lines 169-170, I am not convinced that the meridional curvature at 55-75N, 10 hPa is a good measure of refractive index for stationary planetary waves. The climatological EP fluxes at this latitude band and height location are mainly upward, suggesting the dominant role of vertically propagating Rossby waves. This also implies the important role of the vertical component of the refractive index. It is the first time for me to read that the third term in the equation (5) is highly corrected with meridional curvature term at 55-75N, 10 hPa for the entire winter period from November to March. I would appreciate if the authors can demonstrate the correlation using scatterplots of the reanalysis data also the hindcasts using the 10-day window.

  Thank you for your comment. We revisited the correlations between the second term and third term of the meridional PV gradient, and the second term of the meridional PV gradient with the meridional PV gradient. In Figures 5 to 8 in this document, we show scatter plots of the second and the third term of the meridional PV gradient at 55-75N, 10hPa, and the second term with the meridional PV gradient for the reanalysis data.

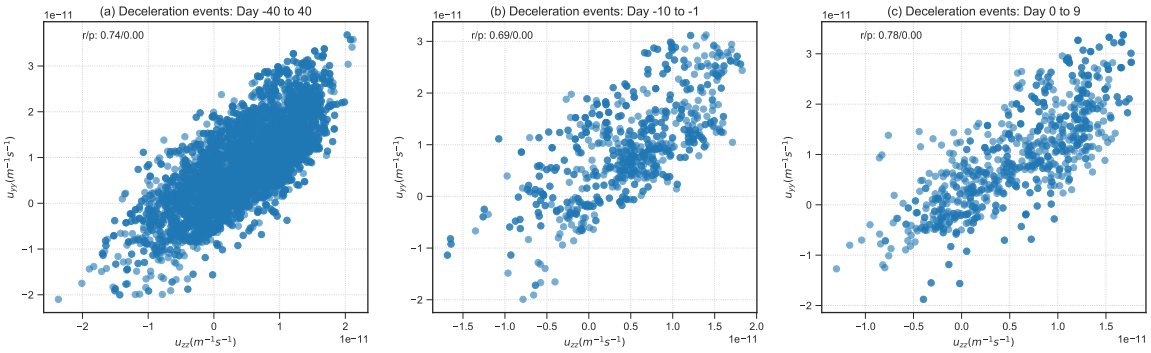

Figure 5: Scatter plots of the daily values of the third term ($\overline{u}_{zz}$) and the second term ($\overline{u}_{yy}$) of the meridional PV gradient for deceleration events during different periods around the events for reanalysis data. Each data point corresponds to a daily value.

In each panel of the scatter plots, we plot the daily values of the corresponding quantity during the specified period. For instance, for days -40 to 40, we plot the daily values of all 81 days against the corresponding quantity of interest on the same day. We compute the Pearson correlation for each scatter plot, the computed coefficient and p-value are included in the legend of each plot.

In Figure 9 in this document, we summarise the computed correlation of all the scatter plots. We show that at 55-75°N, 10 hPa, where $\overline{u}_{yy}$, $\overline{u}_{zz}$, $\overline{q}_y$ are averaged, the second term of the meridional PV gradient correlates well with the third term. The second term of the meridional PV gradient is also highly correlated with the meridional PV gradient. Therefore, we think that for the specific region and level we are looking at, i.e. 55-75°N, 10 hPa, the second term of the meridional PV gradient, $\overline{u}_{yy}$, is a good approximation for the meridional PV gradient and we continue to use $\overline{u}_{yy}$ to approximate the meridional PV gradient.

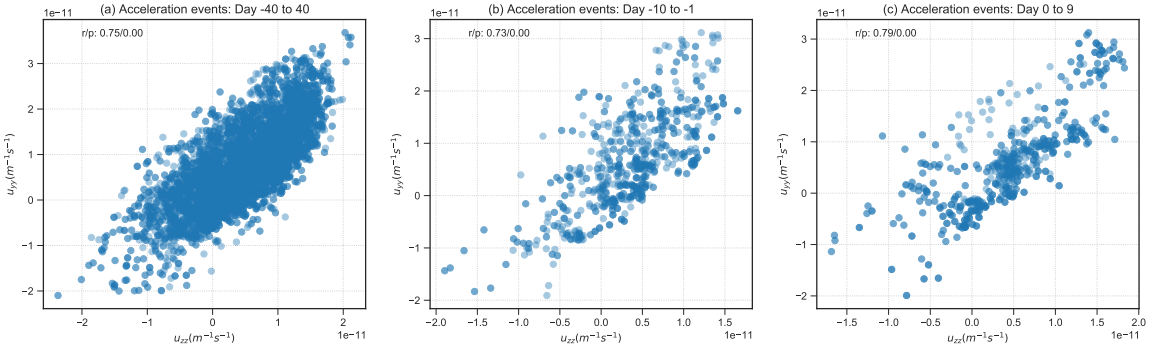

Figure 6: Same as Figure 5 but for acceleration events.

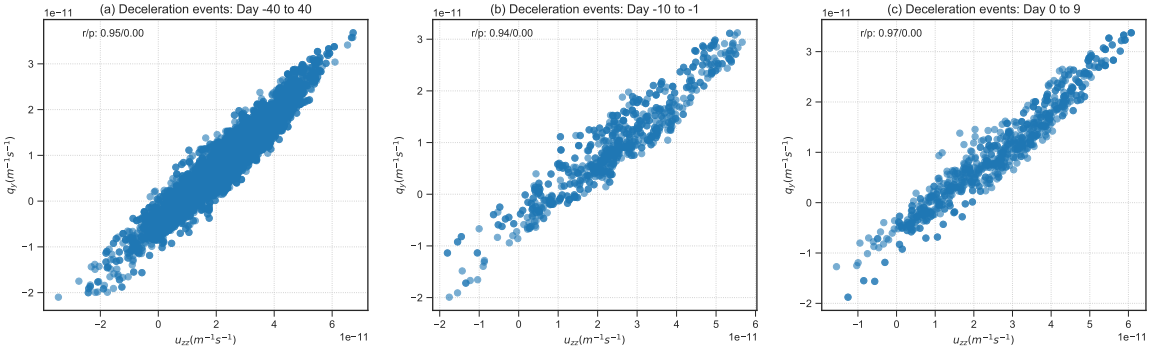

Figure 7: Scatter plots of the daily values of the second term of the meridional PV gradient ($\overline{u}_{yy}$) and the meridional PV gradient ($\overline{q}_y$) for different periods around deceleration events for reanalysis data.

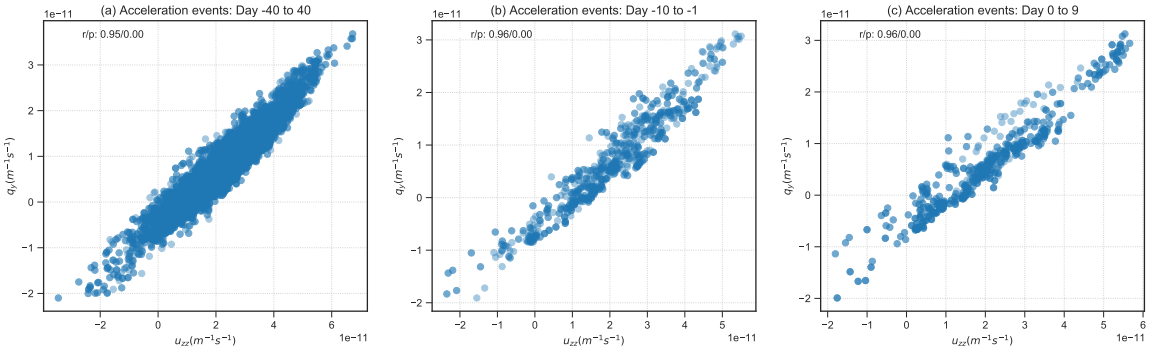

Figure 8: Same as Figure 7 but for acceleration events.

$corr(\overline{u}_{zz}, \overline{u}_{yy})$	-40 to 40	-10 to -1	0 to 9
Deceleration events	0.74/0.00	0.69/0.00	0.78/0.00
Acceleration events	0.75/0.00	0.73/0.00	0.79/0.00

$corr(\overline{q}_y, \overline{u}_{yy})$	-40 to 40	-10 to -1	0 to 9
Deceleration events	0.95/0.00	0.94/0.00	0.97/0.00
Acceleration events	0.95/0.00	0.96/0.00	0.96/0.00

Figure 9: Table summarising the Pearson correlation coefficient (r) and p-value (p) for Figures 5 to 8. Values are written in the format r/p.

- Lines 173-175, are the $\overline{u}_{yy}$ at 55-75N, 10 hPa and v'T' at 45-75N, 100 hPa both calculated using the same 10-day window as well?

  Thank you for pointing this out. In the original manuscript, we take $\overline{u}_{yy}$ at 55-75°N, 10 hPa at day 0 and $\overline{v'T'}$ 45-75°N, 100 hPa integrated over days 0 to 9. However, having considered the comments from both reviewers, we have modified to average $\overline{u}_{yy}$ to be computed over days -10 to -1 instead.

- Figure 3, without losing any information, a, b, and c can be combined into one panel. Also, panels d and e can be combined into one panel.

  Thank you for the suggestion. We have now combined the LTG-30 and 25, LTG-20 and 15, LTG-15 and 5, respectively into a total of 3 panels.

- Figure 4, if all the dashed lines were removed, would it lose any of the key information that the authors want to deliver regarding extreme stratospheric events?

  Thank you for the suggestion. We plotted the dashed lines which refer to the errors corresponding to weak magnitude events in the Figure as a comparison to the strong magnitude events. By comparing the predictability of weak and strong magnitude events, we want to bring out the reason for why the events with stronger magnitude are less predictable. We have also now modified the figure by showing one metric in each panel, comparing the metrics for acceleration and deceleration events of weak and strong magnitude.

- Figure 6, because the focus of the paper is on the predictability, only panels (b) and (e) are worth shown.

  Thank you for the suggestion. As mentioned in Line 292-294 in the manuscript, on average SSWs occur towards the end of the event window, i.e. we identified most SSW events around day 6. Therefore, we still find the panels for lags of days -10 to -1 and days 0 to 9 relevant for extreme events, e.g. SSWs. And as the background state may also matter, we would also need panels (a) and (d), which we have also now modified by taking the day -10 to -1 average for $\overline{u}_{yy}$.

- Figure 7, the temporal evolution of $\overline{u}_{yy}$ is almost identical to that of u itself, this implies that $\overline{u}_{yy}$ at the polar vortex edge is not necessarily a good measure of refractive index as it contains the same high frequency variation that u has.

  Thank you for this comment. In our original manuscript, we were looking at $\overline{u}_{yy}$ at day 0. We have now modified our analysis and take $\overline{u}_{yy}$ averaged over day -10 to -1. We hope the reviewer would agree with us that the 10-day averaged value that is now used in our modified manuscript would help averaging out the high frequency variability contribution.

- Also figure 7, I do not see the LTG-xx lines differ from each other much, what is the purpose of showing them as a multiple panel figure if they can effectively be explained by a sentence or two?

  Thank you for this comment. In Figure 7, we plot the evolution of different variables in different panels, in order compare the evolution of the events identified in the model with the reanalysis events. The lead times are plotted together in the same panel to see if there are any differences in the evolution in the model at different lead times. We have further combined the LTGs following a suggestion by reviewer 1. As we see in Figure 7 that the LTGs lines overlap, this indicates that the mechanisms of the events are well represented in the model for all lead times.

- Figure 8, it is evident that the distributions of uyy have larger variance than those of u. This implies that uyy estimated in this study is not a good measure of waveguide. By definition, a waveguide for stationary Rossby waves should be slow varying. It appears to be measure of PV sharpening on top of background waveguide. Thus, the mechanism that the authors want to study is not captured by uyy.

  Thank you for this comment. In the original manuscript, we use the value of $\overline{u}_{yy}$ at day 0, which is before at the start of the event. As mentioned in earlier replies, we have now modified the $\overline{u}_{yy}$ metric by average day -10 to -1, which we hope to have addressed the problem of high frequency variations. As from the correlation plots shown in Figures 5 to 8 in this document, we find high correlations between $\overline{u}_{yy}$ and $\overline{u}_{zz}$, and between $\overline{u}_{yy}$ and $\overline{q}_y$, respectively. We therefore think for the specific region and level we are looking at, i.e. 55-75°N, 10 hPa, the second term of the meridional PV gradient, $\overline{u}_{yy}$, is a good approximation to the meridional PV gradient and can capture the mechanisms we want to study.

- Figure 9, panels a, b and c of this figure once again suggest that uyy is not a good measure of background waveguide, opposite to what the authors claimed. Its variability is largely associated with wave breaking on both flanks of the polar vortex. Also, this figure be simplified, and the correlations can be summarized by a table or a couple of sentences.

  Thank you for this comment. We think given that $\overline{u}_{yy}$ in the region of investigation has a high correlation with $\overline{q}_y$, $\overline{u}_{yy}$ can be a good measure of the background waveguide. The significant correlation in Figure 9 between the CRPS of $\overline{u}_{yy}$ and $\Delta\overline{u}$ indicates that an improvement in the representation of $\overline{u}_{yy}$ by the model can improve the representation of $\Delta\overline{u}$, though the correlation is not high and thus the influence of $\overline{u}_{yy}$ is not as strong as $\overline{v'T'}$ on predicting $\Delta\overline{u}$. We think the correlation plots can show where each event lies and can point out the two split SSW events which we find might be informative to readers.

References

Birner, T. and Albers, J. R. (2017), 'Sudden Stratospheric Warmings and Anomalous Upward Wave Activity Flux', *SOLA* **13A**(Special Edition), 8–12.

de la Cámara, A., Birner, T. and Albers, J. R. (2019), 'Are Sudden Stratospheric Warmings Preceded by Anomalous Tropospheric Wave Activity?', *J. Clim.* **32**(21), 7173–7189.

Kim, J., Son, S.-W., Gerber, E. P. and Park, H.-S. (2017), 'Defining Sudden Stratospheric Warming in Climate Models: Accounting for Biases in Model Climatologies', *J. Clim.* **30**(14), 5529–5546.

Differences in the Sub-seasonal Predictability of Extreme Stratospheric Events

Rachel W.-Y. Wu[1], Zheng Wu[1], and Daniela I.V. Domeisen[1,2]

[1]ETH Zurich, Zurich, Switzerland
[2]University of Lausanne, Lausanne, Switzerland

Correspondence: Rachel Wai-Ying Wu (rachel.wu@env.ethz.ch)

Abstract. Extreme stratospheric events such as sudden stratospheric warming **(SSW)** and strong vortex events can have downward impacts on surface weather that can last for several weeks to months. Hence, successful predictions of these stratospheric events **can** be beneficial for extended range weather prediction. However, the predictability ~~limit~~ of extreme stratospheric events is most often limited to around 2 weeks or less. The predictability strongly differs **within** events **of the same type**, and **also** between **event types**. The reasons for the observed differences in the predictability, however, are not resolved. **We extend** the **analysis of the predictability of stratospheric extreme events** to **include** wind deceleration and acceleration events**, with SSW and strong vortex events as subsets,** to conduct a systematic comparison of sub-seasonal predictability between **events** in the European Centre for Medium-Range Weather Forecasts (ECMWF) prediction system. **Events of stronger magnitude are found to be less predictable than weaker events for both wind deceleration and acceleration events, with both types of events showing a close to linear dependence of predictability on event magnitude.** There are however deviations from this linear behaviour for very **strong magnitude** events. The difficulties of the prediction system in predicting extremely strong anomalies can be traced to a poor predictability of extreme wave activity **fluxes** in the lower stratosphere, which impacts the prediction of deceleration events, and interestingly, also acceleration events. **Our study suggests that improvements** in the understanding of the wave amplification that is associated with extremely strong wave activity **fluxes** and accurately representing these processes in the model are expected to enhance the predictability of stratospheric extreme events and, by extension, their impacts on surface weather and climate.

1 Introduction

The stratospheric polar vortex (SPV) is a band of strong westerly winds over the polar region at the height of around 20-50km during winter. These circumpolar winds result from a strong temperature gradient in the stratosphere between the polar and subtropical regions during winter due to reduced solar heating over the polar regions. As westerly flow in the stratosphere favours upward wave propagation (Charney and Drazin, 1961), planetary-scale waves formed at the troposphere can propagate

upwards into the stratosphere (e.g. Polvani and Waugh, 2004; Sjoberg and Birner, 2012). Depending on the wave activity and the state of the vortex, the SPV can undergo periods of weakening or strengthening, thus largely varying in strength during the wintertime.

The weakening and strengthening of the SPV can be understood in the framework of wave-mean flow interaction (Matsuno, 1970; Holton and Mass, 1976). Before vortex weakening events, anomalously strong wave activity is observed in the lower stratosphere (Polvani and Waugh, 2004; Hinssen and Ambaum, 2010). The waves can precondition the vortex via wave breaking (Limpasuvan et al., 2004; Albers and Birner, 2014), shaping the vortex structure to be more favourable for upward wave propagation. A preconditioned vortex is associated with a region of large and positive refractive index (Matsuno, 1970; Simpson et al., 2009; Karoly and Hoskins, 1982). As the refractive index for stationary planetary waves is proportional to the meridional potential vorticity (PV) gradient, the meridional PV gradient can be used as a proxy for waveguidability (Albers and Birner, 2014; Jucker and Reichler, 2018). **On the other hand, when wave activity is weak and the SPV is relatively undisturbed, the vortex strengthens on radiative timescales (Limpasuvan et al., 2005; Hitchcock and Shepherd, 2013).** Holton and Mass (1976) demonstrated using a simple mechanistic model that when the wave forcing is below a critical level, the vortex accelerates and approaches a state close to radiative equilibrium.

There exist various definitions to characterise the weak and strong states of the SPV. The most commonly studied events are **major** sudden stratospheric warmings (SSWs, Baldwin et al. (2021)), characterising the abrupt weakening of the SPV. SSW events are commonly defined by the reversal of the SPV mean flow from westerly to easterly (Charlton and Polvani, 2007; Butler et al., 2017; Palmeiro et al., 2015). In some studies, where the primary focus is on the abrupt dynamical nature of SSW events, a definition based on wind change is used (Birner and Albers, 2017; de la Cámara et al., 2019). On the contrary, events where the SPV becomes anomalously strong, with the mean flow accelerating to anomalously strong westerly values beyond a certain threshold, are characterised as strong vortex events (Tripathi et al., 2015). Due to the rapid nature of wave forcing, vortex weakening can be abrupt, whereas **vortex strengthening tends to be** more gradual (Limpasuvan et al., 2005). The more rapid nature and stronger magnitude of vortex weakening than strengthening can be observed by comparing the magnitude of the identified vortex weakening and strengthening events in studies for SPV variability (e.g. Baldwin and Dunkerton, 2001; Limpasuvan et al., 2005). The asymmetry is also observed in the wave activity preceding the events (Polvani and Waugh, 2004) due to the strong relationship between wave forcing and mean flow.

Weak and strong states of the SPV can have a downward impact on surface weather that can last for a few weeks to a few months (Baldwin and Dunkerton, 2001). This downward influence can potentially be used to extend the predictability limit of surface weather from stratospheric origins (Domeisen et al., 2020a). In the stratosphere itself, the deterministic predictability limit of SSW events is about 10 days (Domeisen et al., 2020b; Taguchi, 2020), and it is found that the predictability of SSWs differs strongly between events (e.g. Karpechko, 2018). The source of predictability of SSW events is attributed in some studies to the predictability of wave activity (Stan and Straus, 2009; Karpechko et al., 2018) and tropospheric blocking (e.g. Tripathi et al., 2016), as blocking events often precede SSW events (e.g. Martius et al., 2009). It is found in ensemble forecasting systems that when the forecasts are initialised under strong blocking conditions, ensemble members of the forecasts can undergo bifurcation and lead to large uncertainties (Karpechko, 2018; Lee et al., 2019). However, even when successfully

predicting a preceding blocking event, a model may still fail to predict a SSW (Tripathi et al., 2016), suggesting that **other factors, e.g. the background state of the stratosphere, might be important for successful predictions of SSWs.**

Extreme stratospheric events, e.g. SSW and strong vortex events, are often the main focus of stratospheric predictability studies (e.g. Domeisen et al., 2020b; Taguchi, 2014, 2020). Strong vortex events are shown to be more predictable than SSW events (Domeisen et al., 2020b). To our knowledge, the reason for the observed differences in predictability between event types is, however, not resolved in existing literature, and is often attributed to the different mechanisms driving these events. The sample size of SSW and strong vortex events in sub-seasonal prediction systems tends to be too small to systematically assess their differences in predictability. Thus, in this study, we expand the analysis of the predictability of extreme stratospheric events to wind deceleration and acceleration events . As SSW events and strong vortex events are periods of strong zonal wind deceleration and acceleration, respectively, a better understanding of the predictability of wind deceleration and acceleration events will also contribute to the understanding of the predictability of SSW and strong vortex events. We aim to address the following questions: 1. If we expand the event definitions to wind deceleration and acceleration events, do we also see a difference in predictability between wind deceleration and acceleration events, as for SSW and strong vortex events? 2. If so, what contributes to the difference in predictability between events? For example, is predictability related to event magnitude or event mechanisms? 3. What are the dynamical precursors for the predictability of the events? Do those precursors set the predictability limit of the events?

The paper is structured as follows: Section 2 discusses the data and methods adopted in this study. Section 3.1 illustrates the predictability differences between wind acceleration and deceleration events, Section 3.2 discusses the predictability dependence of events on event magnitude, and Section 3.3 explores the predictability dependence on event mechanisms. Finally, we discuss our results in Section 4.

2 Data and methods

2.1 Datasets

The hindcasts (retrospective forecasts) of the European Centre for Medium-Range Weather Forecasts (ECMWF) model from the subseasonal-to-seasonal (S2S) prediction database (Vitart et al., 2017) are used to evaluate the predictability of stratospheric events in Northern Hemisphere **(NH)** winter, from November to March (NDJFM), in the period of 1998/99-2017/18, which is the full available hindcast period for the model versions used in this study. The hindcasts are initialised twice a week (every Monday and Thursday) for the 20 year period alongside the real-time operational forecasts. The hindcasts consist of 11 ensemble members.

The model versions CY43R3 and CY45R1, corresponding to hindcasts with model version dates of 2017-07-13 to 2019-06-10, are used. Similar model configurations are used in both model versions used here, and they both use the ECMWF ERA-Interim reanalysis (Dee et al., 2011) for initialisation. The different model versions lead to qualitatively similar results in terms of prediction skill in their hindcasts (not shown) and are thus both used for the analysis presented here. The hindcasts are verified against the ERA-Interim reanalysis.

We evaluate the skill of the hindcasts at various lead times. Lead time is referred to as the time between the event onset date and the hindcast initialisation date. For example, a lead time of -5 indicates a hindcast initialised 5 days before the event onset. Hindcasts are divided into 6 lead time groups (LTGs) according to their initialisation dates, each of which represents a 5-day lead time window. For example, LTG-30 refers to hindcasts with initialisation dates of 30 to 26 days before the event, while LTG-5 refers to hindcasts from 5 days to 1 day before the onset date.

2.2 Definition of stratospheric events

From the daily mean of the zonal mean zonal wind at 60° N and 10 hPa ($\overline{u}$) from NDJFM 1998/99-2017/18 of ERA-Interim, we identify zonal wind acceleration and deceleration events. Both acceleration and deceleration events are defined as 10-day events and are identified using a 10-day moving window. Another event can only be identified 20 days after the start of an event to prevent identifying the same event. If a stronger deceleration is observed within 20 days of the last identified event, the period with stronger wind deceleration is selected instead, replacing the weaker event. The start date of the event is defined as day 0 of the event, i.e. the day when acceleration or deceleration starts in the 10-day window. The magnitude of the identified events is defined as the wind change over the 10-day event window, i.e. $\Delta\overline{u} = \overline{u}(t = 9) - \overline{u}(t = 0)$, where t indicates the lead time. **We also impose a criterion that the ratio of the difference between the maximum and minimum wind speed occuring during the 10-day event window has to be less than 1.2, to filter out high frequency variations. Although different processes are involved in deceleration and acceleration events, the duration of wind deceleration and acceleration is found to be similar (Fig. A1a). The 90th percentiles in the duration distributions for both wind deceleration and acceleration are around 10 days. The event magnitude captured by a 10-day window also shows values comparable to the wind changes in SSW and strong vortex events (Fig. A1b). Therefore, after a systematic comparison of different window widths (not shown) and also for comparability between the event types, we use the same event window width of 10 days to identify both wind deceleration and acceleration events.**

To compare the identified acceleration and deceleration events with the extreme stratospheric events, we classify the identified events into weak and strong magnitude events. We choose the 60th percentile of event magnitude as the threshold for strong magnitude events. Events that have an absolute magnitude above the respective 60th percentile of the identified acceleration and deceleration events are classified as strong magnitude events, and those below are classified as weak magnitude events. The 60th percentiles are 16.94 ms^{-1} and -24.55 ms^{-1} for the acceleration and deceleration events, respectively, in the reanalysis. In the ECMWF model, the 60th percentiles of the identified events are 16.77 ms^{-1} and -20.87 ms^{-1} for the acceleration and deceleration events, respectively. The thresholds used here are comparable to the thresholds to define strong deceleration events used in other studies (e.g. Birner and Albers, 2017; de la Cámara et al., 2019). Following Birner and Albers (2017), we compute the standard deviation of deseasonalised daily zonal mean zonal wind and the standard deviation (σ) is found to be around 1 ms^{-1}/day. Our chosen 60th percentile from a 10-day wind change corresponds to daily wind changes of 1.69 ms^{-1} and 2.46 ms^{-1} for strong acceleration and deceleration events, respectively, which corresponds to daily wind changes in the 95th and 99th percentile

Table 1. Identified acceleration and deceleration events from reanalysis. The numbers in the brackets specify the number of events in each category.

Acceleration event	Weak (51)	Strong (34)	Total (85)	**Definition**
Strong vortex	14	11	25	**Following Tripathi et al. (2015) and Domeisen et al. (2020b)**
Vortex recovery	8	11	19	$\overline{u}$ **at any time during event window shows negative values**
Other acceleration events	29	12	41	**-**
Deceleration event	Weak (39)	Strong (26)	Total (65)	**Definition**
SSW	0	10	10	**Following Charlton and Polvani (2007)**
Other deceleration events	39	16	55	**-**

(1.69σ and 2.46σ) in NH Nov-Mar. Thus, the strong magnitude events we define here have magnitudes comparable to SSW and strong vortex events.

For the acceleration and deceleration events identified from reanalysis, we check if they are also associated with extreme stratospheric events, i.e. SSWs, strong vortex events and vortex recovery events. SSW events are defined using the Charlton and Polvani (2007) wind reversal criterion. The onset date of an SSW event is identified as the first day that the daily mean zonal mean zonal winds at 60° N **and** 10 hPa are negative. The winds have to be westerly for at least 20 consecutive days before the event and return to westerly for at least 10 days after the event. We classify a deceleration event to be associated with an SSW event if an SSW occurs within the 10-day event window. The identified deceleration events can also be associated with early final warming (FW) events. Early FW events are defined as in Butler and Domeisen (2021) as those that occur at least 2 days before the median climatological FW date, which is Apr 12 over the period 1979-2019 in JRA-55 reanalysis. Since we only identify events up to March, the number of events associated with final warming events is small, and wave forcing still plays a dominant role in the FW wind reversal. Therefore, we keep the events associated with final warmings in the analysis and do not distinguish them from other deceleration events.

A strong vortex event is defined when $\overline{u}$ exceeds a threshold value. Following Tripathi et al. (2015) and Domeisen et al. (2020b), the chosen threshold value at **60°N and 10 hPa** is 41.2 m/s, which is the 80th percentile of the zonal mean zonal wind averaged from November to March over the 1980-2012 period in ERA-Interim. We classify an acceleration event to be associated with a strong vortex event if the wind at any time during the event window is above this threshold. If the wind at 60° N, 10 hPa at any time during the acceleration event window shows negative wind values, the event is classified as being associated with a vortex recovery event, which occur after SSW events. Table 1 shows the identified events from the reanalysis and their respective event types.

2.3 Skill measures

The following metrics are used to assess the predictability of stratospheric events: Mean error, continuous ranked probability score (CRPS), hit-rate (HR), and ensemble spread. The definitions are stated below.

1. Mean error

   The mean error is the average difference between the hindcast (F) and the observation (O) (here, reanalysis is used instead of observations as the verification dataset). The index i denotes the corresponding ensemble member, and N denotes the ensemble size. For the ECMWF model, N = 11. The perfect score of the mean error is 0.

$$Mean\,Error = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i) \tag{1}$$

2. Continuous ranked probability score (CRPS)

   The CRPS measures the difference between the predicted cumulative distribution function (CDF) ($P_f(x)$) of a variable x and the observed CDF ($P_o(x)$). For ensemble forecasts, the predicted CDF is given by the predictions of all the ensemble members. The perfect score of the CRPS is 0.

   As the CRPS is given by the difference between the predicted and observed distribution, if all ensemble members in a hindcast predict an event magnitude of 0 ms^{-1}, i.e. close to a climatological state where the wind stays relatively constant during a 10-day window, the CRPS of this hindcast will be equal to the observed event magnitude itself.

$$CRPS = \int_{-\infty}^{\infty}(P_f(x) - P_o(x))^2 dx \tag{2}$$

3. Hit-rate (HR)

   The hit-rate (HR) is defined as the fraction of ensemble members that successfully predict an event, given by dividing the number of successful members (M) by the total number of ensemble members (N). A successful prediction requires that the model predicts an event of the same magnitude category as identified from reanalysis, i.e. a strong or weak magnitude event, on the same date as the event in reanalysis. The perfect score of the HR is 1.

$$HR = M/N \tag{3}$$

4. Ensemble spread

The ensemble spread of the ensemble members in a hindcast is measured as the standard deviation of the ensemble member predictions around the ensemble mean ($\overline{F}$). If the ensemble members show perfect agreement with each other, the ensemble spread is 0.

$$Ensemble\,Spread = \sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}(F_i - \overline{F})\right]^2} \tag{4}$$

2.4 Dynamical indices and significance tests

As mentioned in the Introduction, we can quantify the preconditioning of the vortex background state, which guides waves towards the vortex, by the refractive index. As the refractive index is proportional to the meridional PV gradient ($\overline{q}_y$) divided by the zonal mean zonal wind, following Jucker and Reichler (2018) and Albers and Birner (2014), we approximate the refractive index using the meridional PV gradient. Using the formulation of Equation (5) in Simpson et al. (2009), we divide the meridional PV gradient in spherical coordinates ($\overline{q}_\phi$) by the radius of Earth (a) to obtain an equation of the meridional PV gradient in Cartesian coordinates ($\overline{q}_y$),

$$\overline{q}_y = \frac{\overline{q}_\phi}{a} = \frac{2\Omega cos(\phi)}{a} - \left[\frac{(\overline{u}cos\phi)_\phi}{a^2 cos\phi}\right]_\phi + \frac{f^2}{R_d}\left(\frac{p\theta}{T}\frac{\overline{u_p}}{\overline{\theta_p}}\right)_p \tag{5}$$

where ϕ is the latitude, overline denotes the zonal mean, subscripts denote derivatives. As the term associated with Earth's rotation (first term in the equation) is small in extratropical and polar latitudes, and as the third term in the equation correlates well with the second term **in the region we consider, i.e. over 55-75° N at 10 hPa** (not shown), we use the second term in Equation (5), $-\left[\frac{(\overline{u}cos\phi)_\phi}{a^2 cos\phi}\right]_\phi$, as a proxy for waveguidability, hereafter referred to as $\overline{u}_{yy}$. Other than being a reasonable indicator for the refractive index, $\overline{u}_{yy}$ is a measure of the sharpness **of** the edge of the stratospheric polar vortex, thus also a measure of the strength of the vortex state. Similar to Jucker and Reichler (2018), who used a polar cap averaged meridional PV gradient, we take a latitudinal average of $\overline{u}_{yy}$ over 55-75° N at 10 hPa. As a measure of upward wave activity in the lower stratosphere, following Polvani and Waugh (2004), we use the latitudinal average of meridional eddy heat fluxes ($\overline{v'T'}$) over 45-75° N at 100 hPa, where v is the meridional wind, T is the temperature, and prime ($'$) denotes the departure from the zonal mean. **It is, however, important to be aware that the indices $\overline{u}_{yy}$ and $\overline{v'T'}$ might not be independent. To address the interdependency between the indices, we compare correlations between 10-day averaged $\overline{u}_{yy}$ and the 10-day integrated sum of $\overline{v'T'}$ at different time lags. We find the lowest correlation (r = 0.3) between $\overline{u}_{yy}$ averaged over days -10 to -1 with respect to the start date of the stratospheric events and $\overline{v'T'}$ integrated over days 0 to 9 during the events (not shown). Thus, we choose to use the time lags mentioned above to examine the predictability of $\overline{u}_{yy}$ and $\overline{v'T'}$ in the following analyses.**

We use a one-sample t-test to assess the significance for the mean of a distribution. When comparing the significant difference between two distributions, we use a Kolmogorov-Smirnov test (KS test). For both tests, we use a confidence level of 95%. **In**

our analyses, we use linear regression lines as a reference to compare the relationships between event magnitude and precursors, and between their predictability in the model. It is to be noted that we do not intend to imply that the relationships or the dynamics involved are linear.

3 Results

3.1 Predictability of stratospheric events in Northern Hemispheric winter

To illustrate the predictability differences between stratospheric events, we compare the skill of the model in predicting different event types as a function of lead time. The magnitude of the events identified in reanalysis ($\Delta\overline{u}$), measured by the wind difference between day 9 and day 0, predicted by the model hindcasts is compared against the same value in reanalysis for all lead time groups (Fig. 1). The left **panel** in Fig. 1 shows the errors in event magnitude for the deceleration and SSW events (as a subset of deceleration events). The right **panel**, which **has a flipped y-axis**, shows the errors in event magnitude for the acceleration events and strong vortex events (as a subset of acceleration events). Values above the zero line indicate an underestimation of the magnitude of both deceleration and acceleration events, while values below zero indicate an overestimation. The box plots in Fig. 1 of most LTGs lie above zero, indicating an underestimation of event magnitude for both acceleration and deceleration events, including strong vortex events and SSWs. The underestimation of the event magnitude reduces towards smaller LTGs. At LTG-5, the model overestimates around 25% of deceleration and 5% of acceleration events, respectively, shown by the bottom of the box and whisker crossing the zero line. The underestimation of deceleration event magnitude is also seen in Karpechko (2018), where the model shows an initial weakening of the vortex but underestimates the event magnitude.

Previous studies that assessed the predictability of events using event onset dates have found that SSW events are less predictable than strong vortex events (e.g., Domeisen et al., 2020b). This result is confirmed in Fig. 1: The mean errors for SSW events are larger than for strong vortex events, showing that SSW events are less predictable. Extending the analysis to wind deceleration and acceleration events, we also find that deceleration events are associated with larger errors than acceleration events at **all lead times**.

3.2 Predictability dependence on event magnitude

To better understand the nature of the stratospheric events, we plot the distribution of the events identified from reanalysis (transparent bars in Fig. 2, which are the same in all panels). The events identified from reanalysis show an asymmetry in event magnitude, that is, deceleration events are associated with stronger magnitude than acceleration events. The median magnitude **of the wind changes for deceleration and acceleration events in reanalysis** is -21.25 ms^{-1} and 15.32 ms^{-1}, respectively, and -37.22 ms^{-1} and 15.06 ms^{-1}, respectively, for SSW events and strong vortex events. All SSW events belong to the strong deceleration events category, whereas the magnitudes of the strong vortex events are spread more evenly across the weak and strong acceleration event categories (Table 1). The stronger magnitude of deceleration events as compared to acceleration

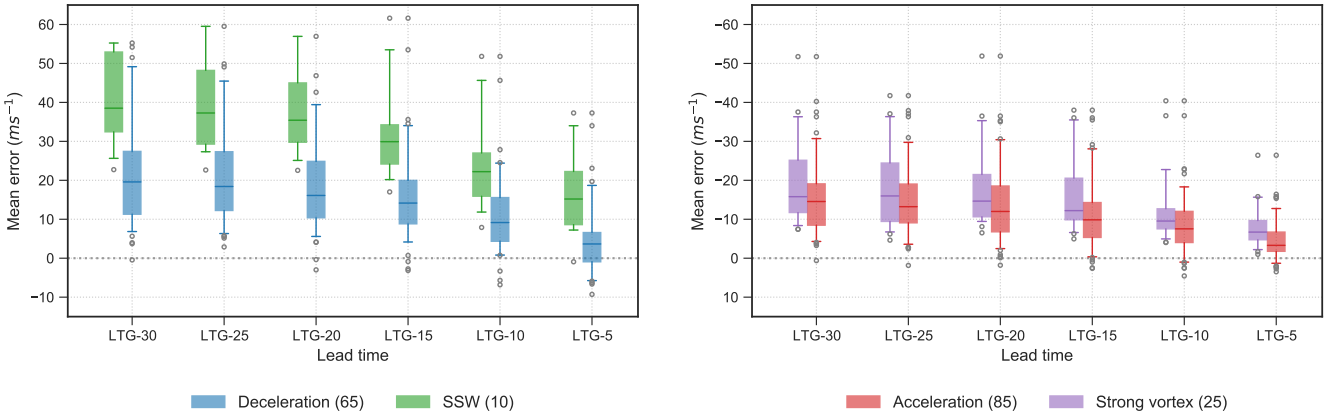

Figure 1. Mean error in the magnitude ($\Delta\overline{u}$) of **(left)** deceleration events (blue), SSW events (green), **and (right)** acceleration events (red) and strong vortex events (purple), for all LTGs. **The y-axis for acceleration events (right panel) is flipped for a more convenient comparison to deceleration events.** The box extends from the 25th to the 75th percentiles of the mean error of the events, with a horizontal line at the median. The whiskers extend from the 5th to the 95th percentiles. Outliers are plotted as grey open circles. The numbers in brackets correspond to the number of events in total for each event type in reanalysis.

events is consistent with Limpasuvan et al. (2005), i.e. that the daily zonal mean zonal wind anomalies observed for vortex weakening events are stronger than for vortex strengthening events.

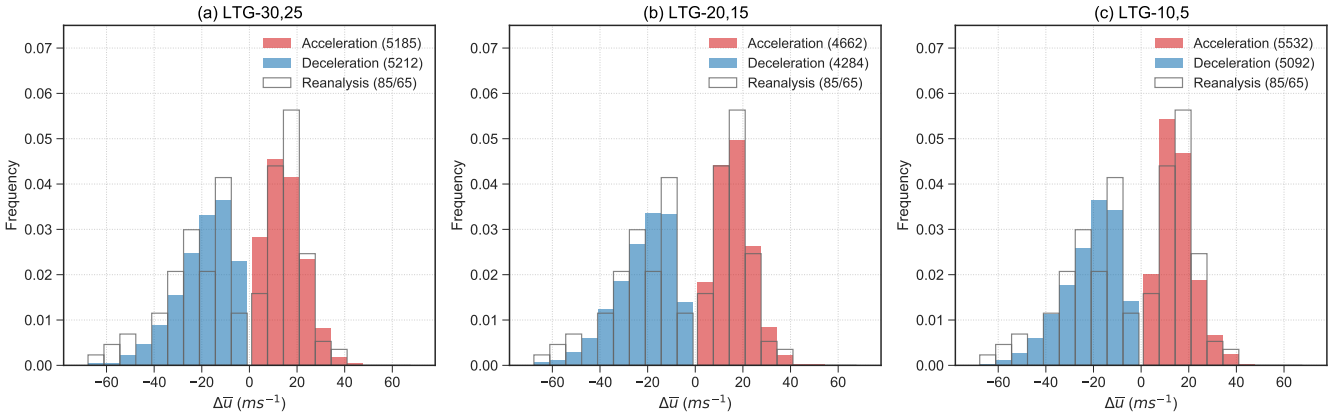

Figure 2. Distributions of the wind change ($\Delta\overline{u}$) of the acceleration (red) and deceleration (blue) events identified from reanalysis (transparent bars with grey outline) and from the ensemble members in the hindcasts that are initialised in NH winter at (a) LTG-30,25, (b) LTG-20,15 and (c) LTG-10,5. Numbers in **parentheses** indicate the number of identified events at each lead time. The reanalysis distributions displayed in all panels are identical and the numbers in brackets refer to the number of acceleration / deceleration events. The histograms are normalised.

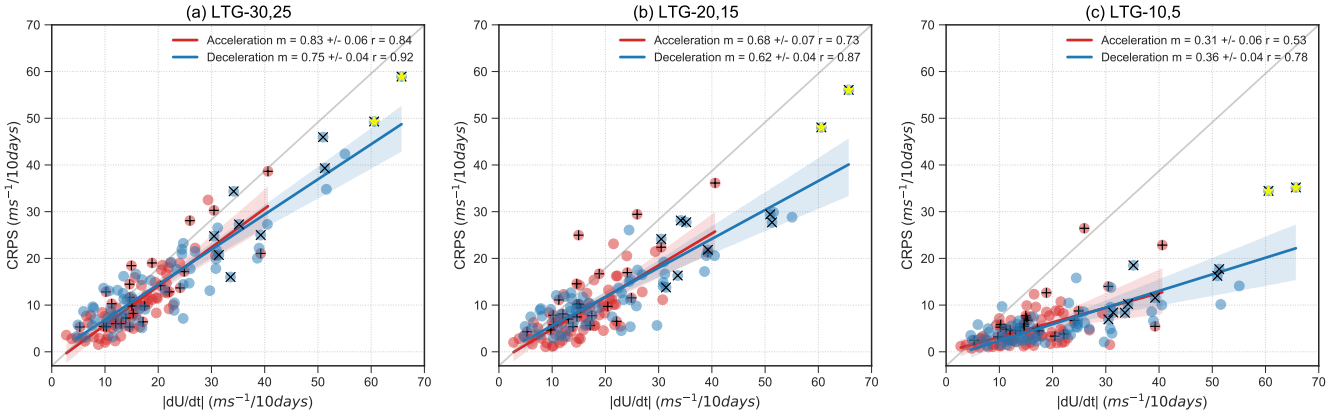

Figure 3. CRPS of event magnitude ($\Delta\overline{u}$) for the identified wind acceleration (red) and deceleration (blue) events plotted against their absolute event magnitude ($|\Delta\overline{u}|$) from reanalysis for different LTGs. The absolute value of event magnitude ($|\Delta\overline{u}|$) is used for a better comparison between acceleration and deceleration events. The filled circles represent strong magnitude events and the empty circles represent weak magnitude events. Linear regression lines are fitted to each of the LTGs, m indicates the slope, including the standard error of the fit. Pearson correlation coefficients (r) are indicated in the legend for acceleration and deceleration events, respectively, and r is statistically significant at 95% for all panels. The shaded region shows the 95% confidence interval of the linear fit. Pluses ('+') indicate events that correspond to strong vortex events and crosses ('×') correspond to SSW events. Yellow stars ('*') denote the 2009 and 2018 split SSW events.

As deceleration events have a stronger magnitude than acceleration events and as the identified events span a wide range of magnitudes, as a first step, we test if the differences in predictability between events arise from different event magnitudes. We plot the CRPS of the model in predicting the event magnitude against the observed event magnitude at different lead times (Fig. 3). **A 1:1 grey diagonal line is added to each panel as a guide to compare the skill of hindcasts to the skill of a climatological prediction (see Section 2.3). Points above the diagonal line show a poorer skill than a climatological prediction, and the points below show a skill that is improved with respect to climatology.** The closer the points are to the x-axis, i.e. the line of CRPS = 0, the more skilful the hindcasts.

For long lead times of around 30 days, the fitted lines lie just below the diagonal line (Fig. 3a), which suggests that the hindcasts exhibit a predictability that is just slightly better than climatological forecasts at these lead times. The fitted slopes then approach the x-axis with decreasing lead time (going from panels (a) to (f)), indicating that, as expected, the model gains more information from initial conditions and the prediction is improved beyond climatological values. The predictability behaviour of both acceleration and deceleration events can roughly be approximated by a linear fit, indicating that the stronger the event magnitude, the less predictable the event. The linear fits corresponding to the deceleration and acceleration events overlap within the 95% confidence interval (blue and red shading, respectively) at all lead times, suggesting that the acceleration and deceleration events show the same predictability behaviour.

At short lead times, most of the points lie close to zero CRPS. **Some events, for instance, the two extreme SSW events with magnitudes of over 60 ms^{-1} (marked by yellow stars in Fig. 3), retain a large CRPS and deviate from the linear fit in the direction of the diagonal line.** The fact that the CRPS remains larger for the two events at LTG-5 suggests that the model might not be capturing the precursors or that it might not accurately represent the mechanisms required to predict these events.

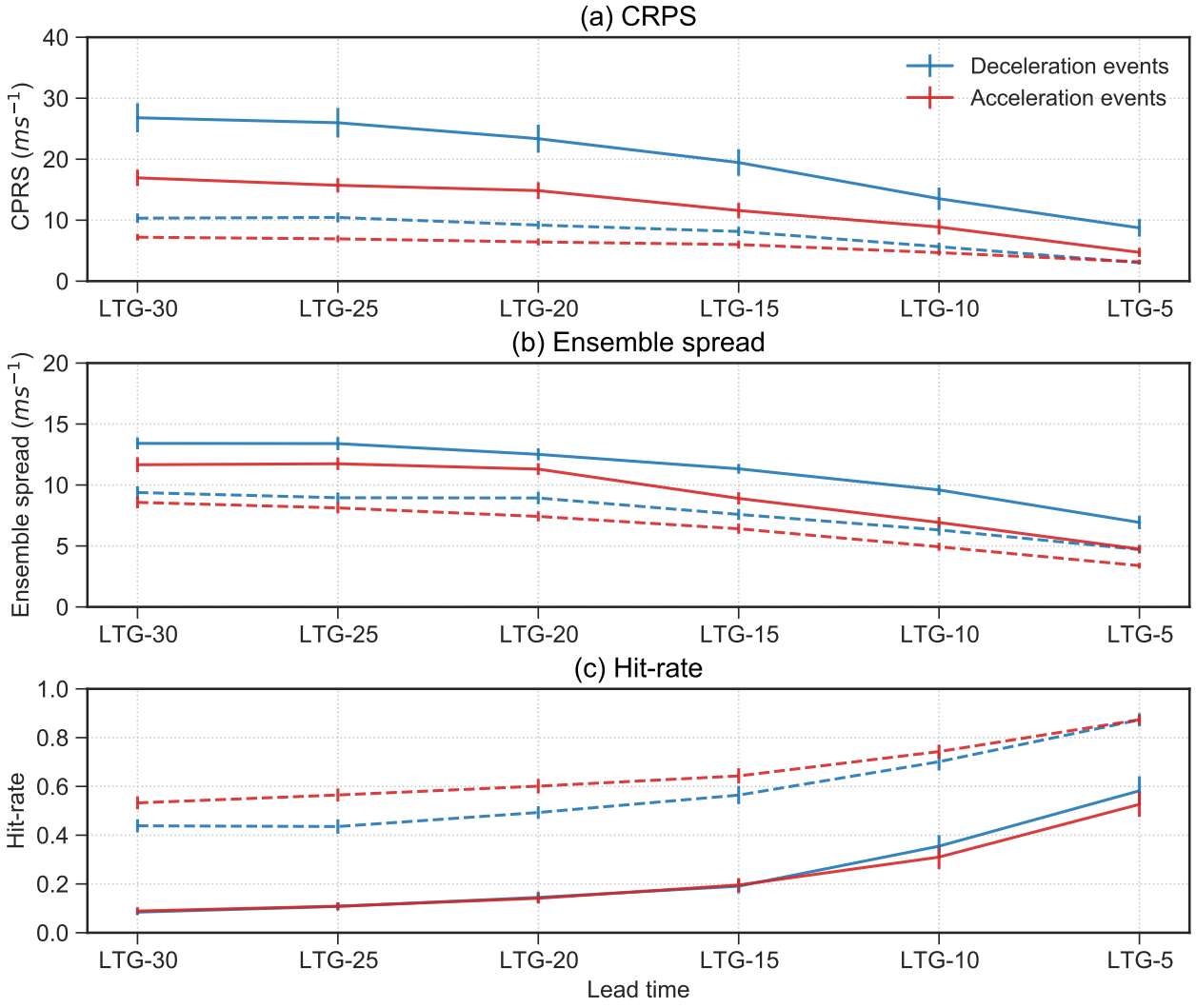

Figure 4. (a) CRPS, (b) ensemble spread and (c) hit-rate for acceleration events (red) and deceleration events (blue) computed by validating the hindcasts against the reanalysis. Solid lines indicate the mean of the strong magnitude events at different LTGs and the dotted lines indicate weak magnitude events. The vertical bars indicate the standard errors for each LTG.

To better illustrate the predictability dependence on event magnitude, we composite the strong and weak magnitude events, i.e. events with magnitudes above and below the 60th percentile, respectively, and compare their averaged skill at different lead times (Fig. 4). Overall, as expected from the model capturing more of the required precursors to predict the events, both acceleration and deceleration events show an increase in hit-rate, and a decrease in ensemble spread and CRPS with decreasing lead time. Strong magnitude events exhibit poorer skill than weak magnitude events, associated with a lower hit-rate, and a larger ensemble spread and CRPS. As large ensemble spread can be observed in ensemble forecasting systems when the forecast is initialised under e.g. strong blocking conditions (Lee et al., 2019; Karpechko, 2018), this might indicate that strong magnitude events are associated with strong precursors or forcings that are not as well captured by the model as those for weak magnitude events. We will discuss the predictability dependence on event mechanism in Section 3.3.

3.3 Predictability dependence on event mechanism

In the last section, we showed that event magnitude strongly determines the predictability, with strong events being less predictable, which can be described mostly by a linear behaviour. Some events, however, deviate from this behaviour, which might be connected the mechanism of the events. In this section, we investigate whether the background state of the SPV and the drivers to the events can have an influence on the predictability of events. We start this section by linking the predictability of the events to the related mechanisms through both the influence of the background state of the stratosphere and the drivers in terms of the upward wave flux for both the reanalysis and the prediction system.

3.3.1 Vortex background state in reanalysis

Before strong deceleration events, $\overline{u}_{yy}$ is significantly stronger than climatology (Fig. 5a), confirming the existing literature that preconditioning of the vortex via sharpening of the vortex edge is often observed before weak vortex events (e.g. Limpasuvan et al., 2004; Jucker and Reichler, 2018). During strong deceleration events (days 0 to 9), $\overline{u}_{yy}$ reduces to a negative value that is significantly weaker than climatology. The vortex recovers at the end of the strong deceleration event and the mean value of $\overline{u}_{yy}$ returns to positive values, but is still significantly weaker than climatology up to 40 days after the event onset. For weak deceleration events, the values of $\overline{u}_{yy}$ before and after the events are close to climatology, and significant signals are only found during the event window (day 0 to 9), suggesting that the preconditioning of the vortex background state before event onset might not be as important for weak deceleration events. For strong acceleration events, increased $\overline{u}_{yy}$ is found only at around 25 days before the events and a few days later, $\overline{u}_{yy}$ decreases to values significantly lower than climatology (Fig. 5c). During strong acceleration events, $\overline{u}_{yy}$ increases to a value significantly above climatology and drifts back to climatology after the event. For weak acceleration events, a few periods of anomalously weak $\overline{u}_{yy}$ are observed at around day -25, 0, 25.

To further illustrate the relationship between $\overline{u}_{yy}$ and event magnitude ($\Delta\overline{u}$), we plot $\Delta\overline{u}$ against $\overline{u}_{yy}$ at **day -10 to -1** for all events (Fig. 6a,d). A significant negative correlation is found between $\overline{u}_{yy}$ and $\Delta\overline{u}$ for deceleration events, that is, the stronger $\overline{u}_{yy}$, the stronger the deceleration. **No significant correlation is found for $\overline{u}_{yy}$ on day -10 to -1 against $\Delta\overline{u}$ for acceleration events (Fig. 6d). The averaged $\overline{u}_{yy}$ for acceleration events, however, shows more negative values than for deceleration**

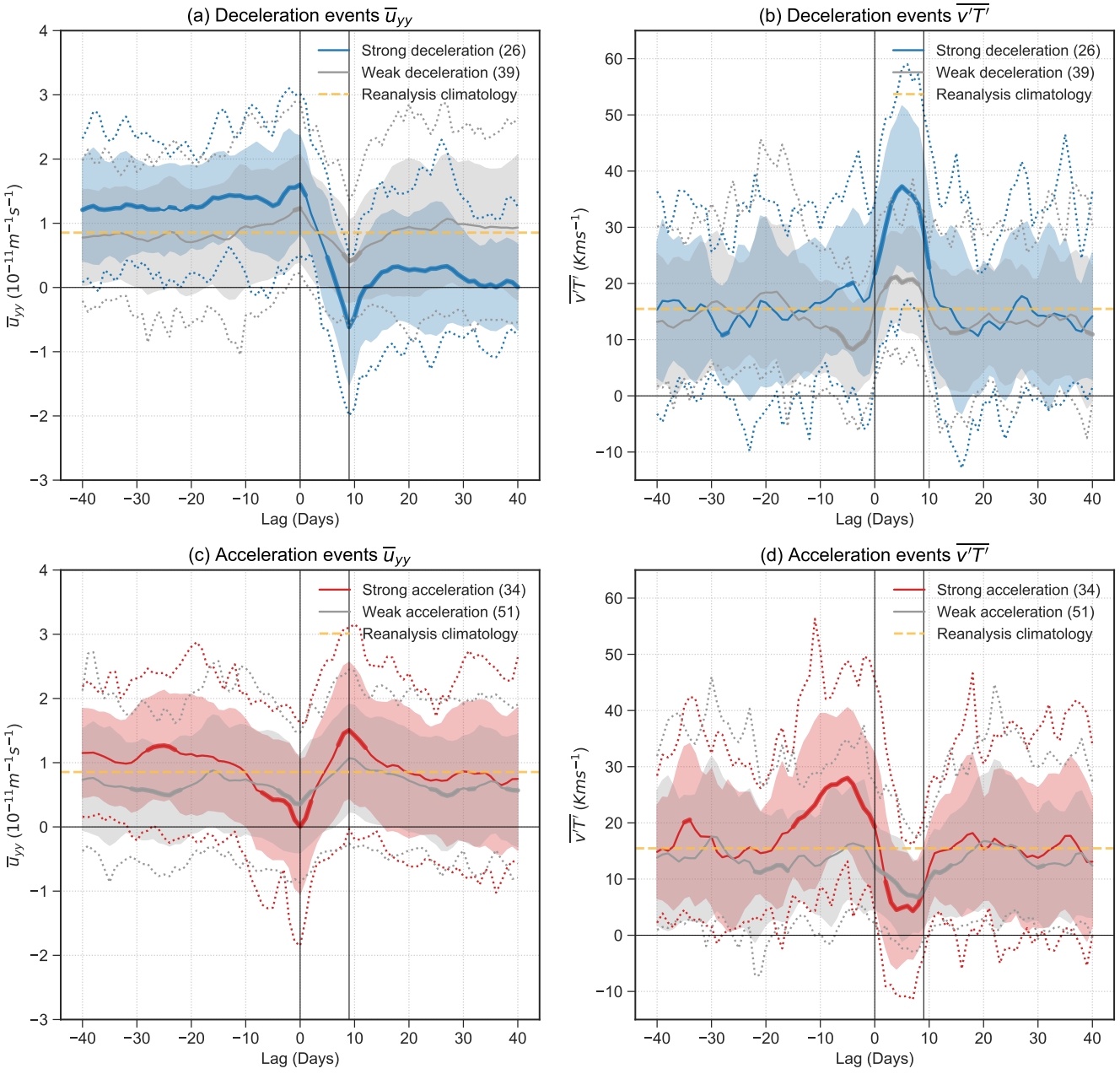

Figure 5. Time evolution of daily values of $\overline{u}_{yy}$ (a, c) and $\overline{v'T'}$ (b, d) for the strong deceleration (blue) (a, b) and acceleration (red) (c, d) events in reanalysis. The solid line is the mean value of all events and the bold parts of the line indicate lags where the composites are significantly different from the reanalysis winter climatology (dotted yellow lines) at 95% using a student's t-test. Weak events are composited separately and shown in grey. The dotted lines in the corresponding colours indicate the 5th and 95th percentiles of the composite, the shaded regions indicate the 25th to 75th percentiles. The numbers in the brackets of the legend indicate the number of events in each composite. Lag is relative to the first day of the identified 10-day events.

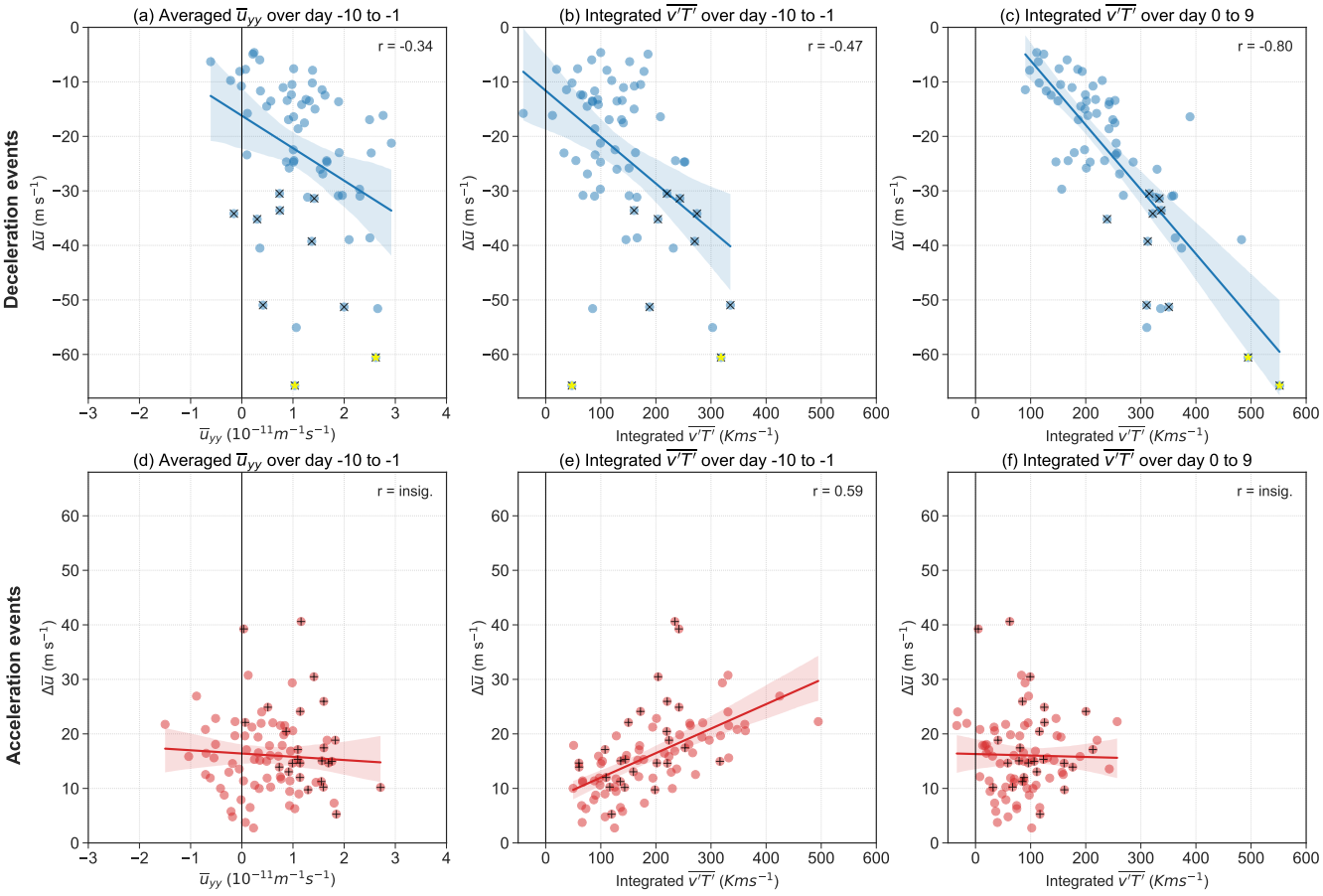

Figure 6. Relationship between the magnitude of the deceleration events ($\Delta\overline{u}$) and (a) $\overline{u}_{yy}$ **for days -10 to -1**, (b) integrated $\overline{v'T'}$ over days -10 to -1 and (c) integrated $\overline{v'T'}$ over days 0 to 9. (d),(e) and (f) same as (a-c) but for the acceleration events. Filled (empty) circles indicate strong (weak) magnitude events, '×' indicates SSW events and '+' indicates strong vortex events. The solid line indicates the fitted linear regression line and shading indicates the 95% confidence interval. Pearson correlation coefficients (r) are significant at 95% in all panels. Yellow stars ('*') in (a-c) denote the 2009 and 2018 split SSW events.

events. Comparing the distributions of the averaged $\overline{u}_{yy}$ for acceleration and deceleration events, the distributions are found to be significantly different from each other (not shown).

3.3.2 Wave activity forcing in reanalysis

285 In addition to the background state, the forcing by drivers is responsible for extreme stratospheric events. In particular, anomalous wave activity in the lower stratosphere drives the deceleration of the SPV mean flow (Polvani and Waugh, 2004; Hinssen and Ambaum, 2010). The 10-day event window captures well the onset of wind deceleration (Fig. A2a) and the anomalously strong wave activity during the event (Fig. 5b). The wave activity starts to increase from day 0, peaks around day 5 and then

decreases to a value that is not significantly different from climatology at the end of the event on day 9. As expected, the wave activity is much stronger during the strong deceleration events than during the weak deceleration events. We find a significant negative correlation when correlating the integrated sum of $\overline{v'T'}$ during the event window (day 0 to 9) with the deceleration event magnitude ($\Delta\overline{u}$) (Fig. 6c). The stronger the wave activity during the event window, the more the wind decelerates. For our definition of deceleration events, using a 10-day event window, SSWs occur on average around day **6** of the event window. Therefore, the peak of $\overline{v'T'}$ during the event window is consistent with our understanding that anomalous wave activity precedes SSW events (e.g. Butler et al. (2017)). A wave activity lower than climatology is found around 10 days before (day -10 to -1) the weak magnitude deceleration events but not the strong magnitude events, suggesting the occurrence of weak acceleration events before weak deceleration events. Wind acceleration is indeed observed 10 days before the weak deceleration events, and the magnitude of the acceleration is similar to the magnitude of the subsequent deceleration events (Fig. A2a). Plotting the integrated $\overline{v'T'}$ for days -10 to -1 against $\Delta\overline{u}$ during the event window, we observe a weak negative correlation (Fig. 6b), which can be explained by low wave activity preceding the weak deceleration events and slightly increased wave activity preceding the strong deceleration events.

Weaker than climatological wave activity $\overline{v'T'}$ is observed during the acceleration event window (Fig. 5d). The wave activity is similar for strong and weak acceleration events but slightly lower for the strong acceleration events. Although strong acceleration events are associated with lower wave activity, there is no significant relationship between the integrated heat flux and event magnitude (Fig. 5f), indicating that wave activity does not drive the acceleration event magnitude, and low wave activity might be more of a threshold criterion for an acceleration event to occur. The passive role of wave activity in wind acceleration is consistent with our understanding that radiative cooling drives the wind acceleration when the wave activity is below a critical level. Interestingly, we observe strong heat flux from around 15 days before the strong acceleration events (Fig. 5d). The same is observed when we exclude vortex recovery events in the composite (not shown). We find a significant positive correlation between the integrated $\overline{v'T'}$ for days -10 to -1 and the wind change over the acceleration event window (day 0 to 9) (Fig. 6e). We find that deceleration events precede about 74% of the strong acceleration events (not shown). The deceleration events that happen before the acceleration events can weaken the vortex, preconditioning the background state of the vortex to be more favourable for the onset of acceleration events, consistent with the weakening of $\overline{u}_{yy}$ before the onset of strong acceleration events from around day -10 (Fig. 5c). The alternation between deceleration and acceleration events is reminiscent of the characteristics of stratospheric vacillations as described in the Holton-Mass model (Holton and Mass, 1976), which shows an oscillation of the mean flow of the vortex after an initial wave forcing.

It is interesting to note that the two strongest strong vortex events, the events with a magnitude of around 40 ms^{-1}, are further away from the linear fit, suggesting that factors other than low wave activity might play a role for these strong magnitude events, for example, strong ozone depletion (e.g., Haase and Matthes, 2019; Lin et al., 2017).

3.3.3 Representation of dynamical processes in the model

Deceleration and acceleration events are found to be driven by anomalies in $\overline{u}_{yy}$ and $\overline{v'T'}$ as described in Section 3.3.1 and 3.3.2 using reanalysis. We now assess the ability of the model to represent these anomalies and relationships. We start by

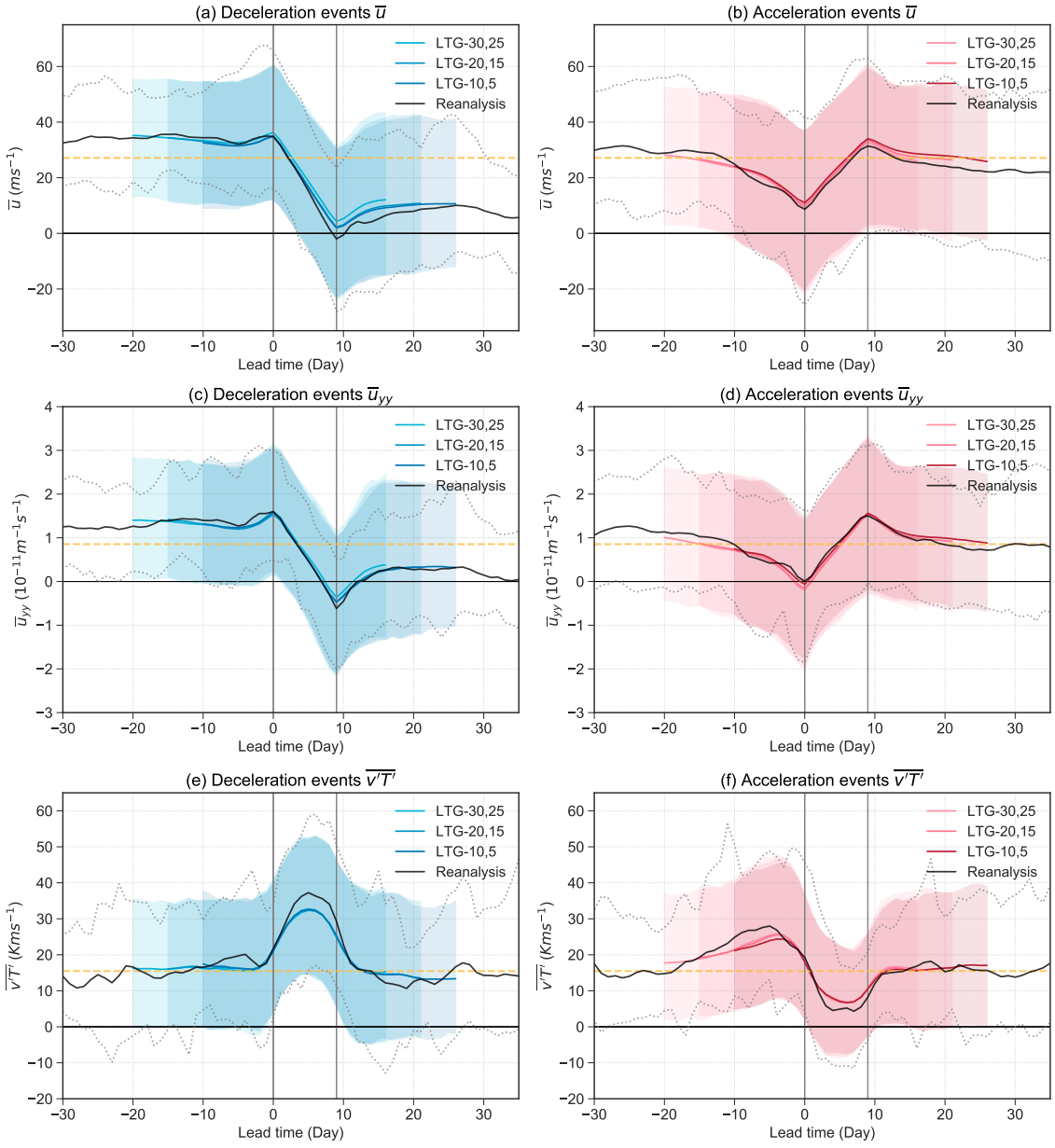

Figure 7. Temporal evolution of $\overline{u}$, $\overline{u}_{yy}$, and $\overline{v'T'}$ for (a),(c),(e) the strong deceleration and (b),(d),(f) the strong acceleration events identified in the model at different LTGs. Solid lines indicate the mean of the event composites and shadings indicate the 5th and 95th percentiles of the events in the prediction system. Black solid lines and black dotted lines indicate the mean, 5th and 95th percentiles for the reanalysis, respectively. Yellow dotted lines show the winter climatology in reanalysis. The first and last 5 days in the mean evolution in the model composite are discarded to account for the different start dates within each LTG.

treating all ensemble members independently and identify deceleration and acceleration events from each separate member at different lead times.

Overall, the climatology of the model event magnitude is similar to that observed in reanalysis at all lead times (Fig. 2). Using a KS test for the model distributions for acceleration and deceleration events, respectively, against the reanalysis distributions, the model and reanalysis distributions are found to not be significantly different from each other. A similar number of events is identified at all lead times, but slightly fewer at LTG-20,15. At LTG-30,25, the model shows an overall underestimation of event magnitude and produces more events with a magnitude close to zero in the model as compared to reanalysis, which is consistent with the predictions being close to climatology at long lead times (Fig. 3a,b). The number of events with very weak magnitude decreases when events are identified at shorter lead times. At all lead times, the model underestimates the number of extremely strong deceleration events (shown by the difference between the model and reanalysis in the negative tails of the distributions). The model covers the range of acceleration event magnitude well but underestimates the frequency of acceleration events with moderate magnitude, i.e. around magnitudes of 20 ms^{-1} over the 10-day event window.

To assess the ability of the model in representing the event mechanisms, we composite the identified strong magnitude events from the model and compare the model evolution of the dynamical variables to the observed evolution in reanalysis (Fig. 7). The model shows a time evolution of $\overline{u}$ similar to that from reanalysis. However, as the model underestimates the number of deceleration events with extremely strong magnitude (Fig. 2), the mean evolution of $\overline{u}$ for deceleration events at all lead times in the model stays above zero, while the winds in reanalysis cross the zero wind line (Fig. 7a). The 5th and 95th percentiles of the model events (shadings) shifted towards more positive $\overline{u}$ as compared to reanalysis (dotted lines) around the end of the event window, indicating that the model does not reach values of $\overline{u}$ that are as small as observed in reanalysis. The mean evolution of $\overline{u}$ at **LTG-30,-25** (lightest blue) remains above the values for all other LTGs throughout the event window until the end of the forecast.

The vortex background state is well represented in the model at all lead times for both strong deceleration and acceleration events. The model shows near identical mean values and variability comparable to the reanalysis for events identified at all lead times (Fig. 7c,d). For the wave forcing (Fig. 7e,f), the model events do not show the extremely high values of $\overline{v'T'}$ during strong deceleration events, or the extremely low values of $\overline{v'T'}$ during strong acceleration events, at all lead times. The 95th percentile of $\overline{v'T'}$ for the deceleration events in reanalysis is outside of the 95th percentile of the model (colour shadings). Similarly, for the acceleration events, the 5th percentile of the reanalysis composite is outside that of the model. Before acceleration events, a peak of $\overline{v'T'}$ is also observed in the model. However, the wave activity in the model for this peak before acceleration events peaks at a later time and at a lower magnitude.

Given the strong relationship observed between event magnitude and wave activity for deceleration events in reanalysis (Fig. 6c), the observed underestimation of strong $\overline{v'T'}$ for deceleration events in the prediction system might explain the observed underestimation of model deceleration event magnitude in Fig. 7a. As a sensitivity experiment, Fig. 7a and 7e are re-plotted by excluding the events with magnitude above the 90th percentile from the reanalysis composite of strong deceleration events (Fig. A3). It is found that the averaged evolution of $\overline{u}$ and $\overline{v'T'}$ of the model composite then covers almost the full variability of the re-computed reanalysis composite, and the evolution of the model composite is near identical to the reanalysis

composite and covers almost the full range of the 5th and 95th percentiles. This suggests that the model has limitations in producing events that have equivalent event magnitudes of above the 90th percentile of the reanalysis deceleration events, likely due to not producing the required strong wave activity. The model also does not show as low $\overline{v'T'}$ during strong acceleration events (Fig. 7f). Although we see the model can produce acceleration events with an evolution similar to reanalysis (Fig. 7b), showing a good variability of acceleration event magnitude in the model, the frequency of acceleration events with moderate event magnitude might still be underestimated (as earlier discussed in Fig. 2). As discussed in Section 3.3.2, $\overline{v'T'}$ might be more a threshold criterion for acceleration events to occur. Specifically, if the wave activity produced in the model is not low enough in some occasions, this can contribute to an underestimation in the number of acceleration events with moderate magnitude, consistent with Figure 2.

We found that overall the model is able to produce events with a range of magnitude similar to reanalysis and has a good representation of event mechanisms. The model, however, has limitations in producing extremely strong anomalies in heat fluxes, thus might be underestimating the number of moderate magnitude acceleration events and the number and magnitude of extremely strong deceleration events. To elucidate the sources of predictability for the events, we now evaluate the magnitude of the anomalies in the precursors, i.e. in $\overline{u}_{yy}$ and $\overline{v'T'}$, captured by the model when predicting the events identified from reanalysis (Fig. 8). **The lead time shown in Fig. 8 (and Fig. 9) is with respect to the start date of the events. Since we are taking the day -10 to -1 average of $\overline{u}_{yy}$ and the lead time is defined to be with respect to the start date of the event, day -10 to -1 is out of the time range of hindcasts with lead time of 10 days or less. Thus, we do not have data and do show plots for LTG-10,5 for $\overline{u}_{yy}$.**

At LTG-30,25, the anomalies in the precursors captured by the model are weak, indicated by the predicted distributions for both acceleration and deceleration events centering around the climatological values (Fig. 8e,h), which is reflected in the fact that acceleration and deceleration events at these lead times are barely separated (Fig. 8a). This is also consistent with Fig. 3 that the points lie close to the diagonal line at long lead times. Nevertheless, the predicted distributions of $\Delta\overline{u}$ are skewed towards the correct signs of the observed events (e.g. the predicted wind change for deceleration events is skewed towards negative values) (Fig. 8a). The predicted distributions for the precursors of acceleration and deceleration events are significantly different from each other, for all lead times, showing that the magnitude of the precursors captured for acceleration and deceleration events are statistically distinguishable even at long lead times.

For shorter lead times, the predicted distributions for acceleration and deceleration become more clearly distinct as the difference between the predicted deceleration and acceleration distribution increases (indicated by greater distance between the distribution means and higher KS test score at shorter lead times). **At LTG-20,15, the model already shows a distribution in $\overline{u}_{yy}$ that is qualitatively similar to reanalysis (Fig. 8f,g), while this is not the case for integrated $\overline{v'T'}$ (Fig. 8i,k) and $\Delta\overline{u}$ (Fig. 8b,d). Even at LTG-10,5**, the model shows a clear underestimation of the very strong deceleration events (with $\Delta\overline{u}$ stronger than -40 ms^{-1}) and an underestimation in the very high values of integrated $\overline{v'T'}$. The frequency of values with around 400 mKs^{-1} are underestimated and values above 400 mKs^{-1} are scarcely predicted in the model (Fig. 8**j,k**). For acceleration events the distribution of the wind change even includes negative values in the predicted event magnitude distribution at LTG-10,5 (Fig. 8c), which is not the case in reanalysis (Figure 8d). As a result, the mean of the predicted $\Delta\overline{u}$ distribution is lower

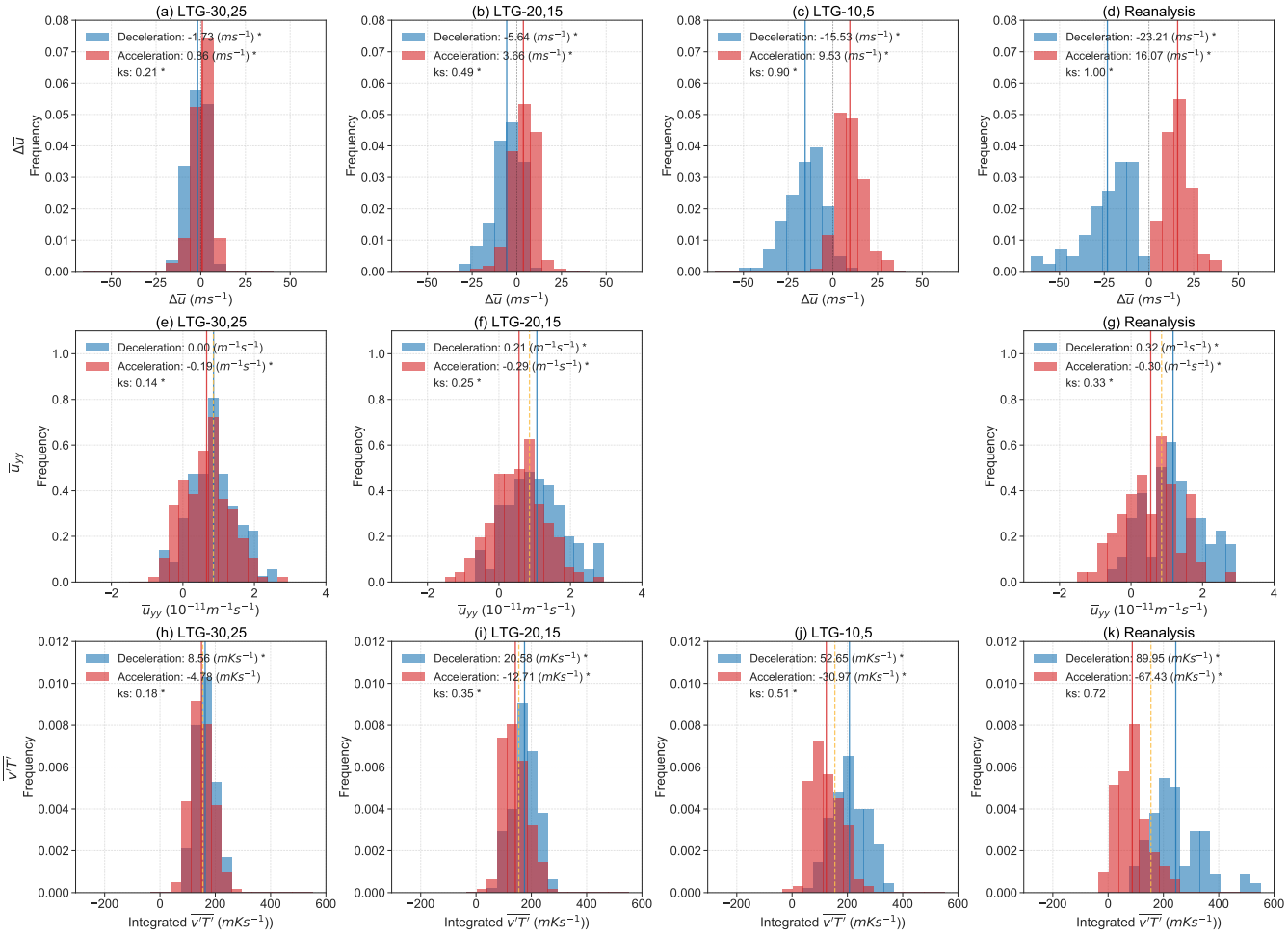

Figure 8. Ensemble mean values of $\Delta\overline{u}$ (a-c), $\overline{u}_{yy}$ **for days -10 to -1** (e,**f**) and integrated $\overline{v'T'}$ over days 0 to 9 (**h-j**) predicted by the model for the events diagnosed in reanalysis at different lead times. The observed distributions from reanalysis are shown in panels (d),(**g**) and (**k**). Deceleration events are shown in blue and acceleration events in red. The winter climatological values of $\overline{u}_{yy}$ and integrated $\overline{v'T'}$ from reanalysis are plotted as yellow dotted lines. The differences of the mean of the distributions from 0 (for $\Delta\overline{u}$) or from climatology (for $\overline{u}_{yy}$ and integrated $\overline{v'T'}$) are shown in the legend. * indicates when the distributions are significantly different from 0 in (a) to (d) and when the distributions differ from the reanalysis climatological value in (e) to (l) using a t-test. The histograms are normalised and a KS test is used to test the significant difference between the acceleration and deceleration event distributions. Statistically significant KS statistics (ks) are indicated by a *.

than for reanalysis. The model predicts more acceleration events with $\Delta\overline{u}$ close to 0 ms^{-1}/ 10 days and some acceleration events with negative $\Delta\overline{u}$. **The model also shows a shift to larger values than that in the reanalysis (Fig. 8j and 8k).**

To quantify the contribution of the predictability of the precursors to the predictability of event magnitude at different lead times, we plot the CRPS of the event magnitude against the CRPS of the precursors (Fig. 9). A significant correlation is

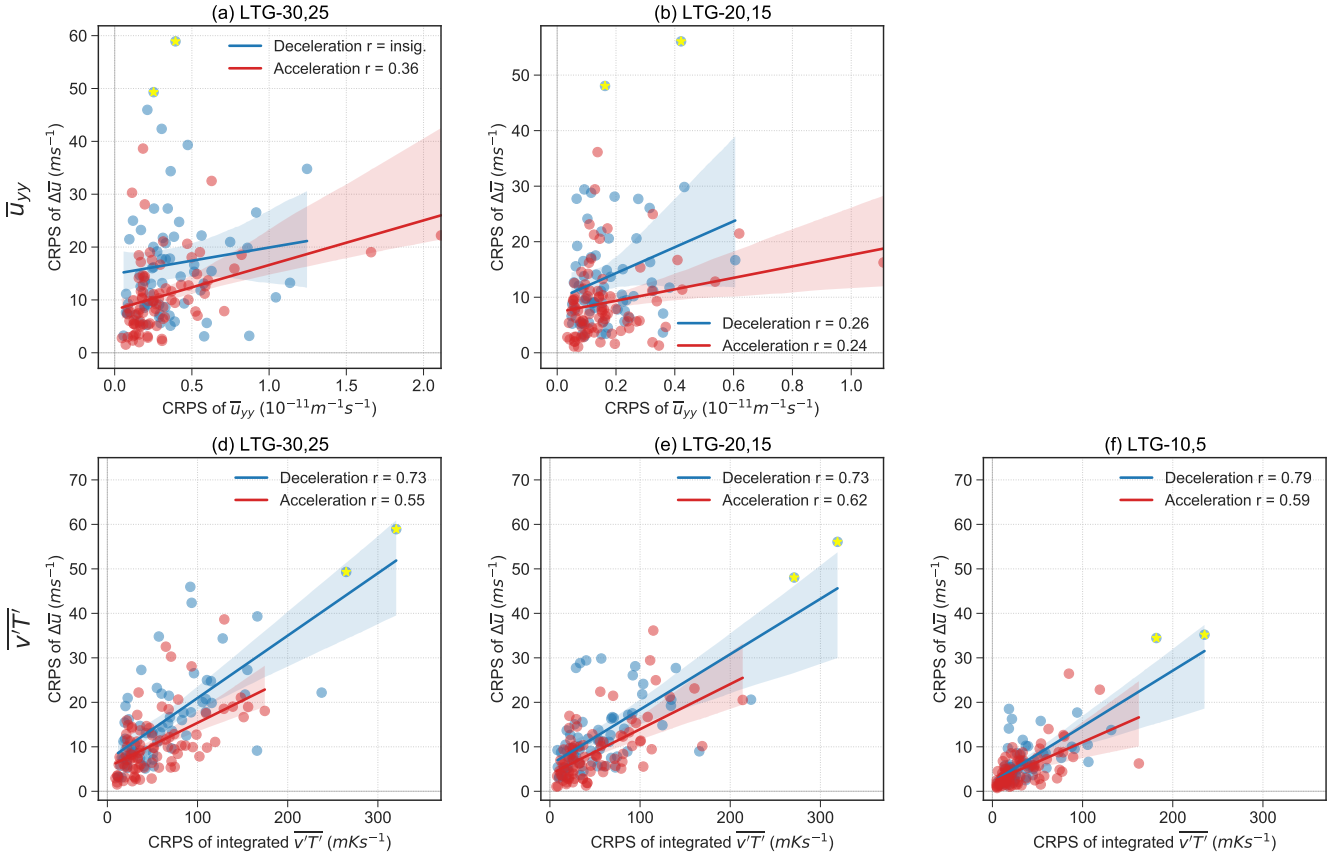

Figure 9. Relationship between the CRPS of $\Delta\overline{u}$ with (a-c) the CRPS of $\overline{u}_{yy}$ on **day -10 to -1** and with (d-f) the CRPS of integrated $\overline{v'T'}$ over days 0 to 9 for the acceleration (red) and deceleration events (blue) identified from reanalysis. The solid line and the shading correspond to the fitted slope and 95% confidence interval of the fit. The Pearson correlation coefficients (r) indicate the correlation in the scatter plots and are statistically significant in all panels at the 95% level. Yellow stars ('*') denote the 2009 and 2018 split SSW events.

found between the CRPS of the event magnitude and of the precursors for both acceleration and deceleration events at all lead times. Consistent with Fig. 8, the model captures the anomalies of the precursors more accurately with decreasing lead time. Specifically, as the CRPS in $\overline{u}_{yy}$ and integrated $\overline{v'T'}$ decreases, the CRPS in $\Delta\overline{u}$ also decreases. On the other hand, the CRPS of $\Delta\overline{u}$ shows a stronger correlation with the CRPS of integrated $\overline{v'T'}$ than $\overline{u}_{yy}$, indicating a stronger contribution of the predictability of integrated $\overline{v'T'}$ to the predictability of event magnitude, which is consistent with the more direct role of $\overline{v'T'}$ than $\overline{u}_{yy}$ in forcing the events.

The largest CRPS values in $\overline{v'T'}$ are found for the 2009 and 2018 split SSW events (yellow stars in Fig. 9). These two SSW events were associated with very strong wave-2 activity (Ayarzagüena et al., 2011; Domeisen et al., 2018). The large $\overline{v'T'}$ for these events might be out of the range of $\overline{v'T'}$ that the model can produce, as suggested earlier, or the mechanisms for these events may not be properly represented in the model. **We further investigate the regional origin of the $\overline{v'T'}$ errors by**

dividing up the regions of heat flux origin into Northern Europe, Siberia, North Pacific and North America/ Greenland (Fig. A4). All regions contribute to the errors, with the largest contribution coming from the North Pacific, and smaller contributions from Northern Europe and Siberia (Fig. A5 and A6). Additional analysis is needed to further understand the origin of the $\overline{v'T'}$ errors.

4 Conclusions

By expanding the stratospheric event definition to wind deceleration and acceleration events using the tendency of the zonal mean zonal wind at 60° N and 10 hPa, we systematically investigate the predictability of extreme events in the SPV in the ECMWF S2S hindcasts. We demonstrate that, overall, the ECMWF model represents the variability of the SPV well in terms of event magnitude and the associated dynamical drivers, and it has a good representation of the dynamical processes that are observed in reanalysis. The model, however, shows limitations in producing events with extremely strong deceleration magnitudes. We find that this is associated with the inability of the model to produce extremely strong wave activity in the lower stratosphere.

The large number of identified deceleration and acceleration events allows us to robustly compare the differences in the event mechanisms in both reanalysis and the model, and to understand the differences in the predictability between events. Consistent with our understanding of the mechanisms of wind deceleration and acceleration events in the framework of wave-mean flow interaction, we find that deceleration and acceleration events are associated with the same anomalies but of opposite signs, namely a strengthened waveguide, in terms of the second meridional derivative of the zonal wind ($\overline{u}_{yy}$), and higher wave activity for deceleration events, measured by the 100 hPa eddy heat flux ($\overline{v'T'}$), and vice versa for acceleration events. The predicted distributions of the acceleration and deceleration events become more distinct at shorter lead times and the respective characteristics of the distributions become better represented. For example, the long tails of deceleration events towards strong events become better represented, although the model continues to underestimate these long tails, even at short lead times.

A large part of the predictability differences between events can be explained by the different event magnitudes. When we express the predictability of deceleration and acceleration events in terms of event magnitude, we found that they both show a predictability dependence on event magnitude; that is, events of stronger magnitude are less predictable. We explain the observed predictability dependence from two perspectives: 1) In a statistical sense, strong magnitude events lie within the tails of the climatological distribution and are penalised more heavily than weak magnitude events, and 2) from a dynamical perspective, strong magnitude events are associated with strong anomalies in $\overline{v'T'}$ and $\overline{u}_{yy}$. The strong precursor anomalies are often less predictable in the model and thus can lead to large uncertainties in event magnitude. The same predictability behaviour with respect to event magnitude for deceleration and acceleration events thus suggests that the observed predictability difference between the event types can to a large extent be explained by the difference in event magnitude between the event types, i.e. the fact that wind deceleration events are associated with greater magnitudes than wind acceleration events, and that SSW events are stronger in magnitude than strong vortex events. We also show that the predictability of the $\overline{v'T'}$ and $\overline{u}_{yy}$ can

explain most of the predictability of the events, with $\overline{v'T'}$ contributing a larger part of the predictability as compared to $\overline{u}_{yy}$. The predictability limit of these dynamical precursors might, therefore, set the predictability limit of events.

For a few events, large errors in the prediction remain even at short lead times, for example, the split SSW events in 2009 and 2018, which are the events with the two strongest event magnitudes of all deceleration and acceleration events investigated in this study. The two split events are reported to be associated with anomalously strong wave-2 wave activity (Harada et al., 2010) and are also reported to be more unpredictable than other SSW events (Rao et al., 2018). The large errors associated with certain events even at short lead time suggest that these events might be associated with mechanisms that are different from weaker magnitude events. For example, internal stratospheric dynamics might play a more important role (e.g. Plumb, 1981; Matthewman and Esler, 2011; Domeisen et al., 2018), which might not be well represented in the model.

Further work is needed to understand the potential reasons as to why the model **has limitations** in producing extremely strong wave activity. **For example, the ability of the model to capture the nonlinear dynamics, which are known to be relevant to SSWs with strong magnitude, has not been explored in this study. These nonlinear processes include the complex behavior of wave breaking, which depending on its exact location and temporal variability can have different effects on the polar vortex, for instance, high frequency wave activity can strengthen the polar vortex rather than weakening it (Harnik, 2009). As such,** one might want to investigate whether the wave amplification mechanisms are different for the 2009 and 2018 split SSW events than for other deceleration events, and to see whether the mechanisms associated with the two events are well represented in the model. A better representation of the wave amplification mechanisms and extremely strong wave activity in the model can potentially enhance the predictability of stratospheric events, and by extension their impacts on surface weather and climate.

Appendix A

To choose a suitable event window width for identifying acceleration and deceleration events, we study the variability of the SPV through the tendency in the zonal mean zonal wind at 60° N, 10 hPa in reanalysis. We identify periods that show consecutive days of wind acceleration and deceleration by counting the number of consecutive days that the daily wind change is of the same sign. If the wind changes sign on one day, that day is counted as a new period of wind change. The number of days in the identified wind change period is defined as the duration.

The duration distribution for wind acceleration and deceleration is qualitatively similar to each other, both following an exponential distribution (Fig. A1a). The magnitude is given by the wind change over the duration of a given identified period. The duration and event magnitude shows a near relationship (Fig. A1b).

As an extra analysis, we have separated the reanalysis data of meridional heat flux into contributions from four regions that are selected based on the existing literature and on our own analysis. Specifically, based on the meridional heat flux composite of deceleration events (Fig. A4a), we divided the 45-75°N latitude region equally into four regions as indicated in Figure A4: (a) Northern Europe (40°W - 50°E), (b) Siberia (50°E - 140°E), (c) North Pacific (140°E - 130°W) and (d) North America/ Greenland (130°W - 40°W). The composites with respect to day 0 to 9 of the decelera-

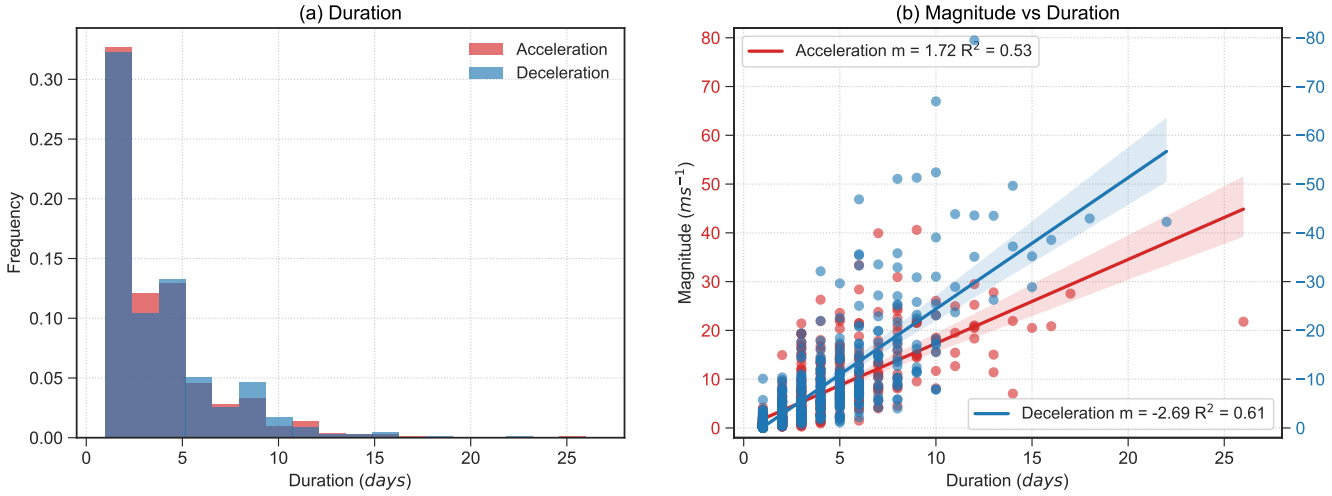

Figure A1. Periods of wind acceleration (red) and deceleration (blue) in reanalysis. (a) Duration, (b) the relationship between the magnitude and duration of the wind acceleration and deceleration periods. For acceleration periods (red), refer to the red axis on the left. For deceleration periods (blue), refer to the blue axis on the right. The solid lines mark the linear fit to the scatter plots and the shading marks the 95% confidence interval of the fit. The histograms are normalised.

tion events in comparison to the November to March average show anomalously positive heat flux in the three regions, namely, Northern Europe, Siberia and North Pacific, and anomalously negative averaged heat flux in the region North America/ Greenland. The composite for acceleration events shows similar patterns. Thus, we choose to average the heat
475 **flux over the same four regions for both deceleration and acceleration events to examine the predictability of the wave activity captured by the model at different lead times.**

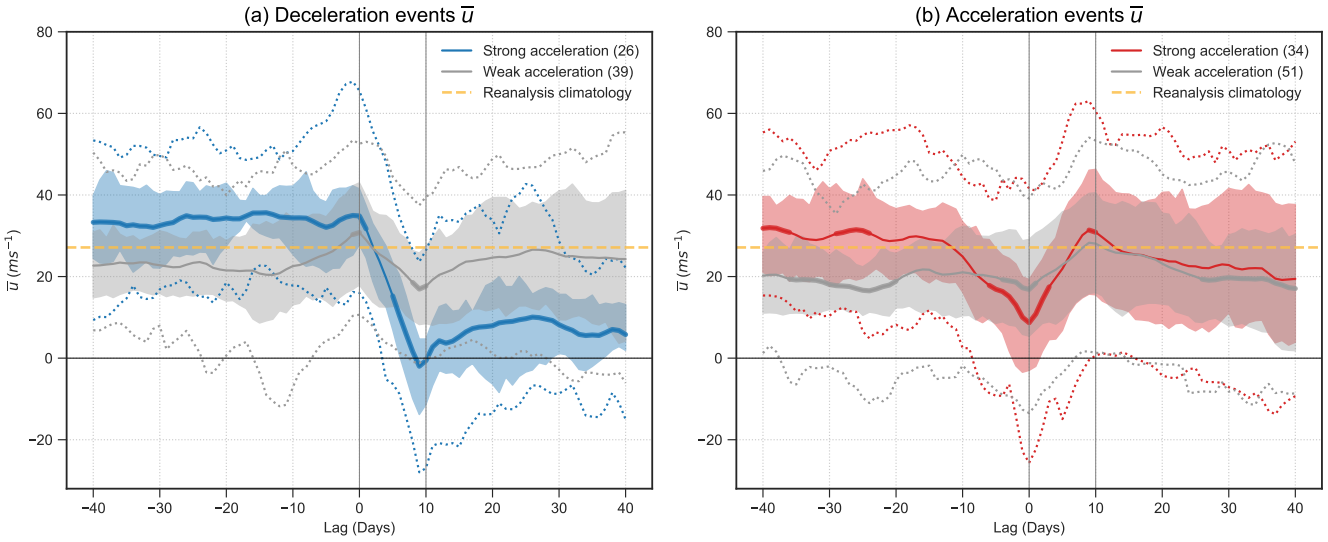

Figure A2. Time evolution of daily values of $\overline{u}$ for the strong deceleration (blue) and acceleration (red) events in reanalysis. The solid line shows the mean value and the bold line indicates where the composites are significantly different from the reanalysis climatology using a student's t-test. Weak events are composited separately and shown in grey. The dotted lines in the corresponding colours indicate the 5th and 95th percentile of the composite, the shaded region indicates the 25th to 75th percentiles. The dotted yellow line shows the winter climatology $\overline{u}$ in reanalysis. The number in the brackets of the legend indicates the number of events in the composites. Lag is relative to the first day of the identified 10-day events.

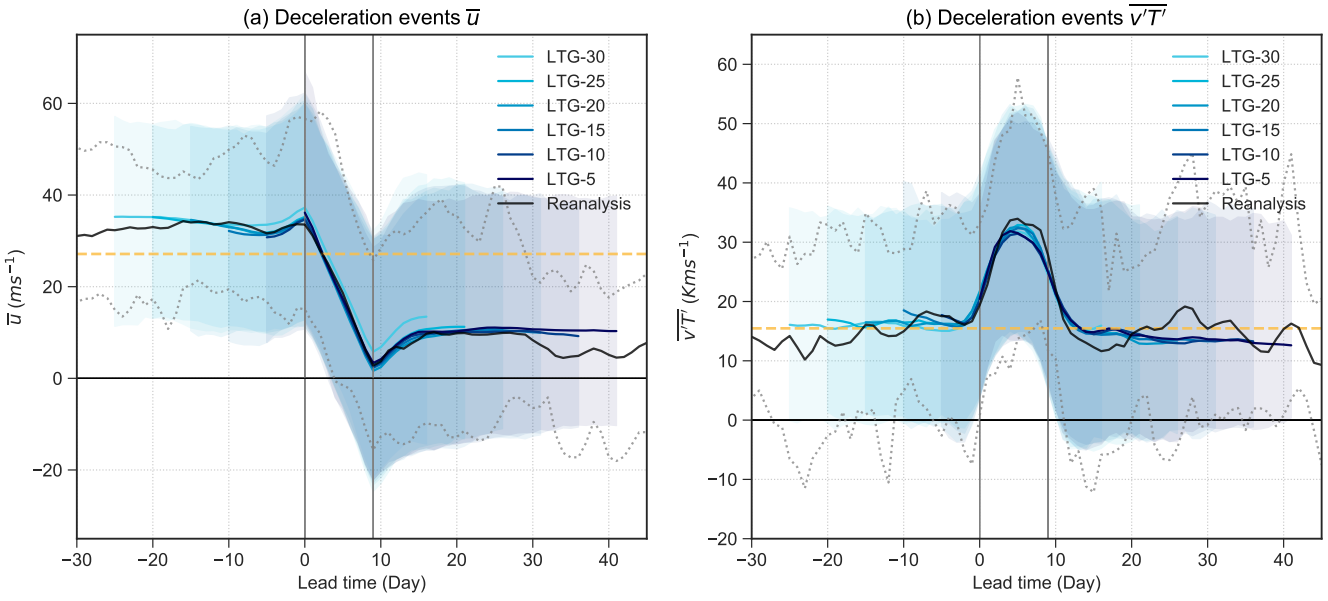

Figure A3. Like Figures 7a and 7e but excluding events with magnitude above the 90th percentile in the reanalysis composite.

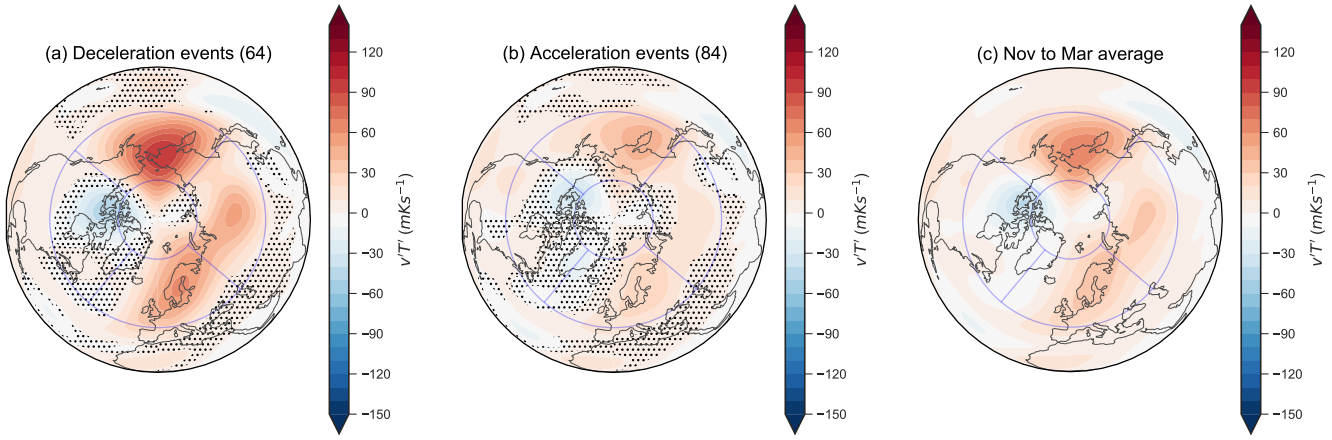

Figure A4. Composite $v'T'$ at 100 hPa averaged over day 0 to 9 (during the event window) of (a) Deceleration events, (b) Acceleration events and (c) Nov to Mar average in reanalysis. Blue lines mark the regions of investigation: Northern Europe (40°W - 50°E), Siberia (50°E - 140°E), North Pacific (140°E - 130°W) and North America / Greenland (130°W - 40°W). Numbers in parentheses indicate the number of events in the composite and unhatched regions in (a) and (b) indicate areas found to be significantly different from (c) using a t-test.

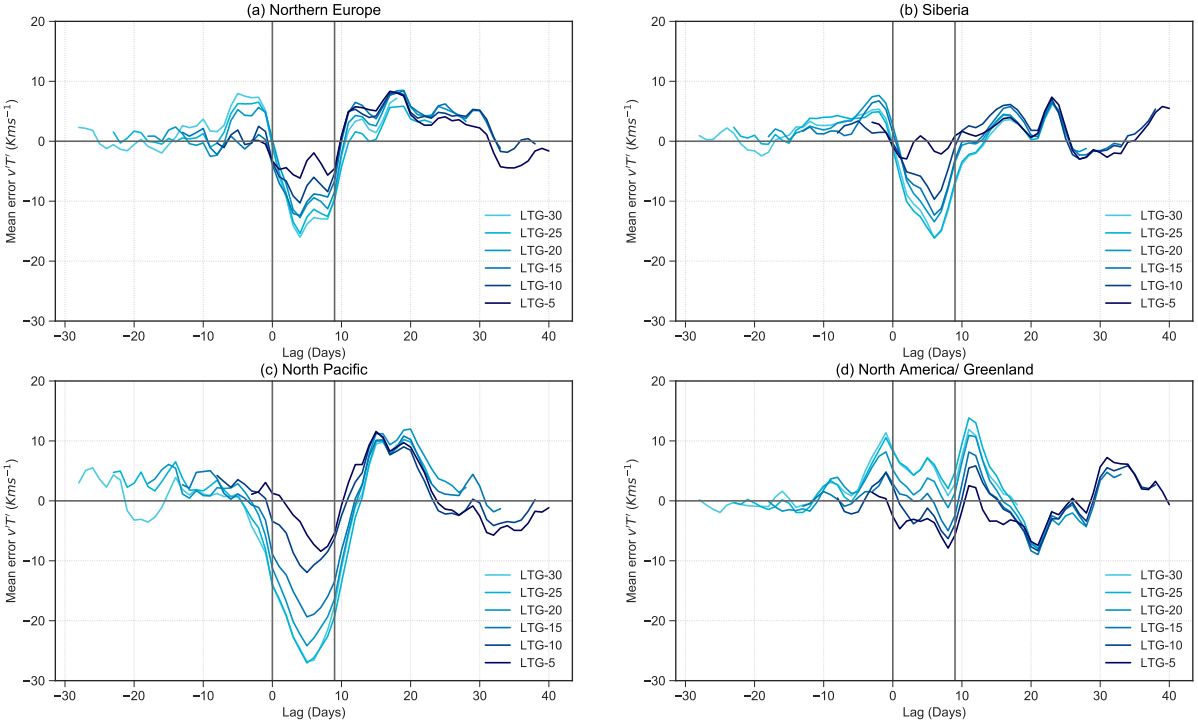

Figure A5. Mean error of the composite $v'T'$ at 100 hPa for deceleration events over (a) Northern Europe, (b) Siberia, (c) North Pacific and (d) North America / Greenland predicted by the hindcasts at different lead times.

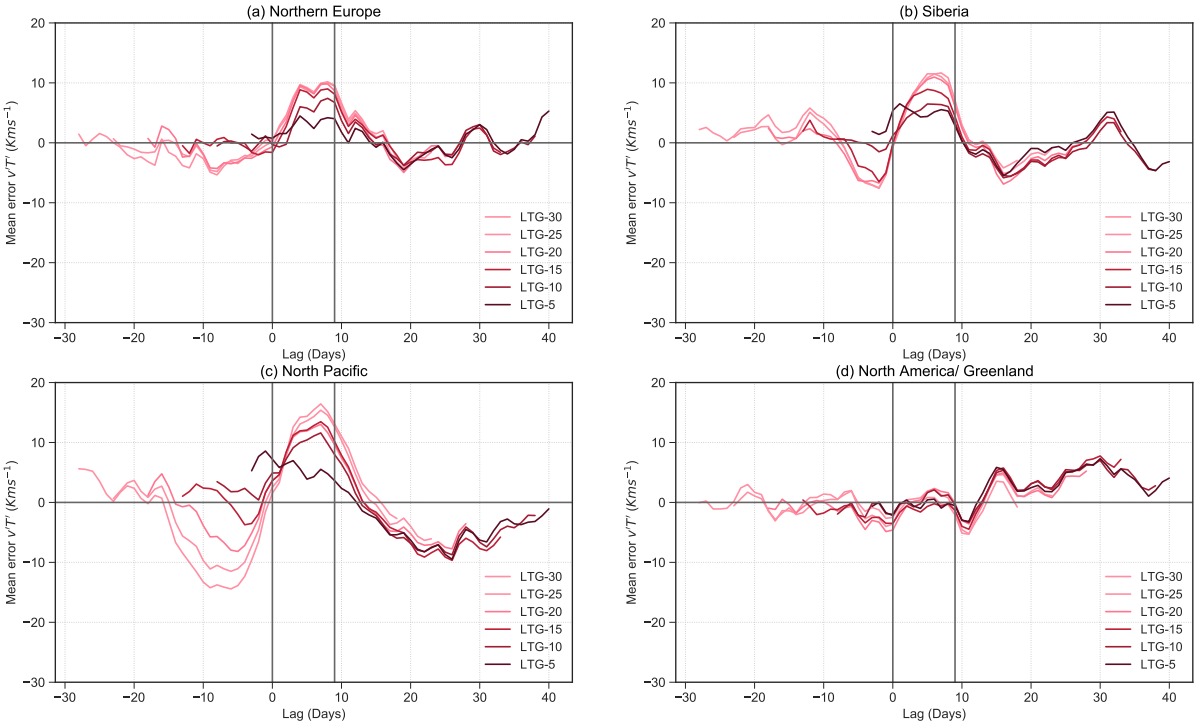

Figure A6. Same as Fig. A5 but for acceleration events.

Author contributions. R.W. and D.D. designed the study. R.W. performed the analysis, made the figures, and wrote the manuscript draft. All authors discussed the research and worked on revising the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. The authors thank Huw Davies, Hilla Afargan-Gerstman, and Bernat Jiménez-Esteve for helpful discussions on earlier versions of this research, and Ole Wulff for help with the S2S data. The work of R.W. is funded through ETH grant ETH-05 19-1 "How predictable are sudden stratospheric warming events?". The work of Z.W. is partially funded by the Swiss Data Science Center within the project *EXPECT* (C18-08). Support from the Swiss National Science Foundation through projects PP00P2_170523 to Z.W. and D.D. and PP00P2_198896 to D.D. is gratefully acknowledged. The ERA-Interim data was obtained from https://apps.ecmwf.int/datasets/ and the S2S data was obtained from https://apps.ecmwf.int/datasets/data/s2s-reforecasts-instantaneous-accum-ecmf/levtype=pl/type=cf/.

References

Albers, J. R. and Birner, T.: Vortex preconditioning due to planetary and gravity waves prior to sudden stratospheric warmings, Journal of the Atmospheric Sciences, 71, 4028–4054, https://doi.org/10.1175/JAS-D-14-0026.1, 2014.

Ayarzagüena, B., Langematz, U., and Serrano, E.: Tropospheric forcing of the stratosphere: A comparative study of the two different major stratospheric warmings in 2009 and 2010, J. Geophys. Res. Atmos., 116, https://doi.org/10.1029/2010JD015023, 2011.

Baldwin, M. P. and Dunkerton, T. J.: Stratospheric Harbingers of Anomalous Weather Regimes, Science, 294, https://doi.org/10.1126/science.1063315, 2001.

Baldwin, M. P., Ayarzagüena, B., Birner, T., Butchart, N., Butler, A. H., Charlton-Perez, A. J., Domeisen, D. I. V., Garfinkel, C. I., Garny, H., Gerber, E. P., Hegglin, M. I., Langematz, U., and Pedatella, N. M.: Sudden Stratospheric Warmings, Rev. Geophys., 59, e2020RG000 708, https://doi.org/10.1029/2020RG000708, 2021.

Birner, T. and Albers, J. R.: Sudden Stratospheric Warmings and Anomalous Upward Wave Activity Flux, SOLA, 13A, 8–12, https://doi.org/10.2151/sola.13A-002, 2017.

Butler, A. H. and Domeisen, D. I. V.: The wave geometry of final stratospheric warming events, Weather Clim. Dyn., 2, 453–474, https://doi.org/10.5194/wcd-2-453-2021, 2021.

Butler, A. H., Sjoberg, J. P., Seidel, D. J., and Rosenlof, K. H.: A sudden stratospheric warming compendium, Earth Syst. Sci. Data, 9, 63–76, https://doi.org/10.5194/essd-9-63-2017, 2017.

Charlton, A. J. and Polvani, L. M.: A New Look at Stratospheric Sudden Warmings. Part I: Climatology and Modeling Benchmarks, Journal of Climate, 20, 449–469, 2007.

Charney, J. G. and Drazin, P. G.: Propagation of planetary-scale disturbances from the lower into the upper atmosphere, J. Geophys. Res., 66, 83–109, https://doi.org/10.1029/JZ066i001p00083, 1961.

de la Cámara, A., Birner, T., and Albers, J. R.: Are Sudden Stratospheric Warmings Preceded by Anomalous Tropospheric Wave Activity?, J. Clim., 32, 7173–7189, https://doi.org/10.1175/JCLI-D-19-0269.1, 2019.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorolog. Soc., 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E. P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S. W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 2. Predictability Arising From Stratosphere-Troposphere Coupling, Journal of Geophysical Research: Atmospheres, 125, https://doi.org/10.1029/2019JD030923, 2020a.

Domeisen, D. I., Butler, A. H., Charlton-Perez, A. J., Ayarzagüena, B., Baldwin, M. P., Dunn-Sigouin, E., Furtado, J. C., Garfinkel, C. I., Hitchcock, P., Karpechko, A. Y., Kim, H., Knight, J., Lang, A. L., Lim, E. P., Marshall, A., Roff, G., Schwartz, C., Simpson, I. R., Son, S. W., and Taguchi, M.: The Role of the Stratosphere in Subseasonal to Seasonal Prediction: 1. Predictability of the Stratosphere, Journal of Geophysical Research: Atmospheres, 125, https://doi.org/10.1029/2019JD030920, 2020b.

Domeisen, D. I. V., Martius, O., and Esteve, B. J.: Rossby Wave Propagation into the Northern Hemisphere Stratosphere: The Role of Zonal Phase Speed, Geophysical Research Letters, 45, 2064–2071, 2018.

Haase, S. and Matthes, K.: The importance of interactive chemistry for stratosphere-troposphere coupling, Atmospheric Chemistry and Physics, 19, 3417–3432, https://doi.org/10.5194/acp-19-3417-2019, 2019.

Harada, Y., Goto, A., Hasegawa, H., Fujikawa, N., Naoe, H., and Hirooka, T.: A Major Stratospheric Sudden Warming Event in January 2009, J. Atmos. Sci., 67, 2052–2069, https://doi.org/10.1175/2009JAS3320.1, 2010.

Harnik, N.: Observed stratospheric downward reflection and its relation to upward pulses of wave activity, J. Geophys. Res. Atmos., 114, https://doi.org/10.1029/2008JD010493, 2009.

Hinssen, Y. B. and Ambaum, M. H.: Relation between the 100-hPa heat flux and stratospheric potential vorticity, Journal of the Atmospheric Sciences, 67, 4017–4027, https://doi.org/10.1175/2010JAS3569.1, 2010.

Hitchcock, P. and Shepherd, T. G.: Zonal-Mean Dynamics of Extended Recoveries from Stratospheric Sudden Warmings, J. Atmos. Sci., 70, 688–707, https://doi.org/10.1175/JAS-D-12-0111.1, 2013.

Holton, J. R. and Mass, C.: Stratospheric Vacillation Cycles, J. Atmos. Sci., 33, 2218–2225, https://doi.org/10.1175/1520-0469(1976)033<2218:SVC>2.0.CO;2, 1976.

Jucker, M. and Reichler, T.: Dynamical Precursors for Statistical Prediction of Stratospheric Sudden Warming Events, Geophysical Research Letters, 45, 124–13, https://doi.org/10.1029/2018GL080691, 2018.

Karoly, D. J. and Hoskins, B. J.: Three Dimensional Propagation of Planetary Waves, Journal of the Meteorological Society of Japan. Ser. II, 60, 109–123, https://doi.org/10.2151/jmsj1965.60.1_109, 1982.

Karpechko, A. Y.: Predictability of sudden stratospheric warmings in the ECMWF extended-range forecast system, Monthly Weather Review, 146, 1063–1075, https://doi.org/10.1175/MWR-D-17-0317.1, 2018.

Karpechko, A. Y., Charlton-Perez, A., Balmaseda, M., Tyrrell, N., and Vitart, F.: Predicting Sudden Stratospheric Warming 2018 and Its Climate Impacts With a Multimodel Ensemble, Geophysical Research Letters, 45, 513–538, https://doi.org/10.1029/2018GL081091, 2018.

Lee, S. H., Charlton-Perez, A. J., Furtado, J. C., and Woolnough, S. J.: Abrupt Stratospheric Vortex Weakening Associated With North Atlantic Anticyclonic Wave Breaking, Journal of Geophysical Research: Atmospheres, 124, 8563–8575, https://doi.org/10.1029/2019JD030940, 2019.

Limpasuvan, V., Thompson, D. W. J., and Hartmann, D. L.: The Life Cycle of the Northern Hemisphere Sudden Stratospheric Warmings, J. Clim., 17, 2584–2596, https://doi.org/10.1175/1520-0442(2004)017<2584:TLCOTN>2.0.CO;2, 2004.

Limpasuvan, V., Hartmann, D. L., Thompson, D. W. J., Jeev, K., and Yung, Y. L.: Stratosphere-troposphere evolution during polar vortex intensification, J. Geophys. Res. Atmos., 110, https://doi.org/10.1029/2005JD006302, 2005.

Lin, P., Paynter, D., Polvani, L., Correa, G. J. P., Ming, Y., and Ramaswamy, V.: Dependence of model-simulated response to ozone depletion on stratospheric polar vortex climatology, Geophysical Research Letters, 44, 6391–6398, https://doi.org/https://doi.org/10.1002/2017GL073862, 2017.

Martius, O., Polvani, L. M., and Davies, H. C.: Blocking precursors to stratospheric sudden warming events, Geophysical Research Letters, 36, https://doi.org/10.1029/2009GL038776, 2009.

Matsuno, T.: Vertical Propagation of Stationary Planetary Waves in the Winter Northern Hemisphere, J. Atmos. Sci., 27, 871–883, https://doi.org/10.1175/1520-0469(1970)027<0871:VPOSPW>2.0.CO;2, 1970.

Matthewman, N. J. and Esler, J. G.: Stratospheric sudden warmings as self-tuning resonances. Part I: Vortex splitting events, Journal of the Atmospheric Sciences, 68, 2481–2504, https://doi.org/10.1175/JAS-D-11-07.1, 2011.

Palmeiro, F. M., Barriopedro, D., García-Herrera, R., and Calvo, N.: Comparing Sudden Stratospheric Warming Definitions in Reanalysis Data, J. Clim., 28, 6823–6840, https://doi.org/10.1175/JCLI-D-15-0004.1, 2015.

Plumb, R. A.: Instability of the distorted polar night vortex: A theory of stratospheric warmings, Journal of the Atmospheric Sciences, 38, 2514–2531, https://doi.org/10.1175/1520-0469(1981)038<2514:IOTDPN>2.0.CO;2, 1981.

Polvani, L. M. and Waugh, D. W.: Upward Wave Activity Flux as a Precursor to Extreme Stratospheric Events and Subsequent Anomalous Surface Weather Regimes, J. Clim., 17, 3548–3554, https://doi.org/10.1175/1520-0442(2004)017<3548:UWAFAA>2.0.CO;2, 2004.

Rao, J., Ren, R., Chen, H., Yu, Y., and Zhou, Y.: The Stratospheric Sudden Warming Event in February 2018 and its Prediction by a Climate System Model, J. Geophys. Res. Atmos., 123, 13,332–13,345, https://doi.org/10.1029/2018JD028908, 2018.

Simpson, I. R., Blackburn, M., and Haigh, J. D.: The role of eddies in driving the tropospheric response to stratospheric heating perturbations, Journal of the Atmospheric Sciences, 66, 1347–1365, https://doi.org/10.1175/2008JAS2758.1, 2009.

Sjoberg, J. P. and Birner, T.: Transient tropospheric forcing of sudden stratospheric warmings, Journal of the Atmospheric Sciences, 69, 3420–3432, https://doi.org/10.1175/JAS-D-11-0195.1, 2012.

Stan, C. and Straus, D. M.: Stratospheric predictability and sudden stratospheric warming events, J. Geophys. Res. Atmos., 114, https://doi.org/10.1029/2008JD011277, 2009.

Taguchi, M.: Predictability of major stratospheric sudden warmings of the vortex split type: Case study of the 2002 southern event and the 2009 and 1989 northern events, Journal of the Atmospheric Sciences, 71, 2886–2904, https://doi.org/10.1175/JAS-D-13-078.1, 2014.

Taguchi, M.: Verification of Subseasonal-to-Seasonal Forecasts for Major Stratospheric Sudden Warmings in Northern Winter from 1998/99 to 2012/13, Advances in Atmospheric Sciences, 37, 250–258, https://doi.org/10.1007/s00376-019-9195-6, 2020.

Tripathi, O. P., Charlton-Perez, A., Sigmond, M., and Vitart, F.: Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions, Environ. Res. Lett., 10, 104 007, https://doi.org/10.1088/1748-9326/10/10/104007, 2015.

Tripathi, O. P., Baldwin, M., Charlton-Perez, A., Charron, M., Cheung, J. C. H., Eckermann, S. D., Gerber, E., Jackson, D. R., Kuroda, Y., Lang, A., McLay, J., Mizuta, R., Reynolds, C., Roff, G., Sigmond, M., Son, S. W., and Stockdale, T.: Examining the Predictability of the Stratospheric Sudden Warming of January 2013 Using Multiple NWP Systems, Monthly Weather Review, 144, 1935–1960, https://doi.org/10.1175/MWR-D-15-0010.1, 2016.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H. S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D. J., Xiao, H., Zaripov, R., and Zhang, L.: The subseasonal to seasonal (S2S) prediction project database, Bulletin of the American Meteorological Society, 98, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1, 2017.