Response to Reviewers

We would like to thank both reviewers for their time and effort to review our manuscript. We are happy that Reviewer 1 acknowledges our effort in developing statistical tests to detect significant and robust signals despite the sparse sampling size. We thank Reviewer 2 for the constructive criticism. The reviewer comments will help to balance the discussion of the various aspects related to stratosphere-troposphere coupling around SSW events and therefore to provide a contribution by adding a tropospheric viewpoint to this longstanding discussion.

Our intention was and is to use tropospheric weather regimes to stress the variety of the tropospheric response, and in particular that the tropospheric response can be divided into major categories that also tend to influence the subsequent tropospheric evolution. Our intention is also to raise even more awareness that the canonical response of NAO negative conditions and a cold surge in Europe, as indicated from composite studies, can often be misleading. We concur that due to our wording the initial version implies too much causality. We also concur that if we had intended to show causality we would have to discuss the underlying mechanisms explaining why the tropospheric state of the NAE sector affects the aftermath of an SSW in the troposphere. Instead, with the study at hand we aim to shed light on the potential role of the troposphere in the response to stratospheric variability and to suggest weather regimes at SSW onset as a potential indicator of the subsequent tropospheric impact. This does not exclude other processes but might be an important additional factor.

In order to make this intention of our study clearer and to achieve a more balanced discussion of our results, we will rewrite our study and thereby accommodate the reviewer comments. In a revised manuscript we will highlight the potential role of the troposphere in stratosphere-troposphere coupling following SSW events. In particular, we will more carefully introduce the tropospheric state in the NAE sector in terms of the refined 7 weather regimes as a potential factor which preconditions the impact of SSWs.

Answer to major reviewer comments

In the following we comment on the major points raised by both reviewers and how we aim to address these. This response will be followed by a detailed response to all reviewer comments and a revised manuscript.

1. **Extending the analysis to S2S prediction models:** We fully agree that extending the analysis to S2S prediction models will be extremely worthwhile. However, it is currently unknown to what extent complex prediction models are indeed able to represent the variety of tropospheric responses to stratospheric forcing. Although simplified models indeed show a role of the troposphere in the downward impact, as reviewer 2 points out, this has not been sufficiently tested in more complex models beyond the canonical response and selected case studies. From a preliminary

analysis of S2S prediction model data we anticipate large biases and a very complex role of the representation of stratosphere - troposphere coupling in prediction models. This will be complex to disentangle, and we will therefore not be able to cover the analysis of the model data in this study. We will comment on this in the revised version of the manuscript.

2. **Role of the lower stratospheric persistence in the downward response:** We agree that the persistence of lower stratospheric anomalies is a very important issue, and we will include a more thorough discussion of this matter and the relevant citations in the manuscript. An inital analysis revealed that 2 out of 5 SSW events with the GL regime at the onset have a persistent lower stratospheric temperature response, while 4 out of 6 EuBl cases have a persistent lower stratospheric temperature response. This goes in the right direction by indicating that the shorter (longer) persistence in the lower stratosphere for the SSWs associated with GL (EuBl) may add some support to the persistence of the tropospheric response, but the statistics are too small to provide a clear result. We will expand on this in more detail in the new manuscript.

3. **Statistical testing:** We fully agree that the number of SSWs in the observational record of the satellite era is very small. Thus, we have put significant effort into the design of the statistical tests, as also commended by Reviewer 1, to take sampling uncertainty into account and obtain meaningful results. We recognize that in the manuscript this procedure may not have been explained in sufficient detail. We will expand on this in the revised version of the manuscript. Furthermore, we clarify the procedure again in the following.

The overarching question we address in this study is whether after SSWs we can detect robust tropospheric geopotential height anomalies and whether these anomalies are significantly different from situations without an SSW. Hence, the relevant null hypothesis is that the tropospheric geopotential height anomalies after SSWs are indistinguishable from geopotential height anomalies commonly occurring in the absence of an SSW. The testing procedure follows two important steps:

1. First, we assess the **robustness** of the samples by performing a Monte Carlo resampling. For that purpose, the dripping paint plots are re-computed by resampling the original samples 100 times with repetitions. This yields confidence intervals of the dripping paint plots, estimating the uncertainty inherent in each sample. Due to the small sample size, these confidence intervals are relatively large.

2. Second, we compute 1000 random samples of the same size as the original sample but for random periods with the same weather regime(s) at the central date but no SSW occurring within ± 60 days, yielding estimates of the distributions of geopotential height anomalies occurring in the absence of SSWs. Testing for **significance** is done by comparing the confidence intervals and distributions obtained from the random samples.

Following this procedure, we thus show that the anomalies in the ALL composite are not highly robust (Fig. A4a in the manuscript), indicative of the large variability in the tropospheric response.

Yet, the anomalies observed between 10-20 days after the SSW are statistically significantly different from non-SSW periods. Now, we can ask the question whether there are subsamples of all SSWs that show a more robust response that is also statistically different from periods with no SSW. We select these subsamples according to the weather regime present at the time of the SSW. Note that the null hypothesis here is still the same as for the ALL sample, namely that no significantly different anomalies occur. Thus, we find that particularly robust and significant anomalies occur if the SSW is timed with European blocking, for example.

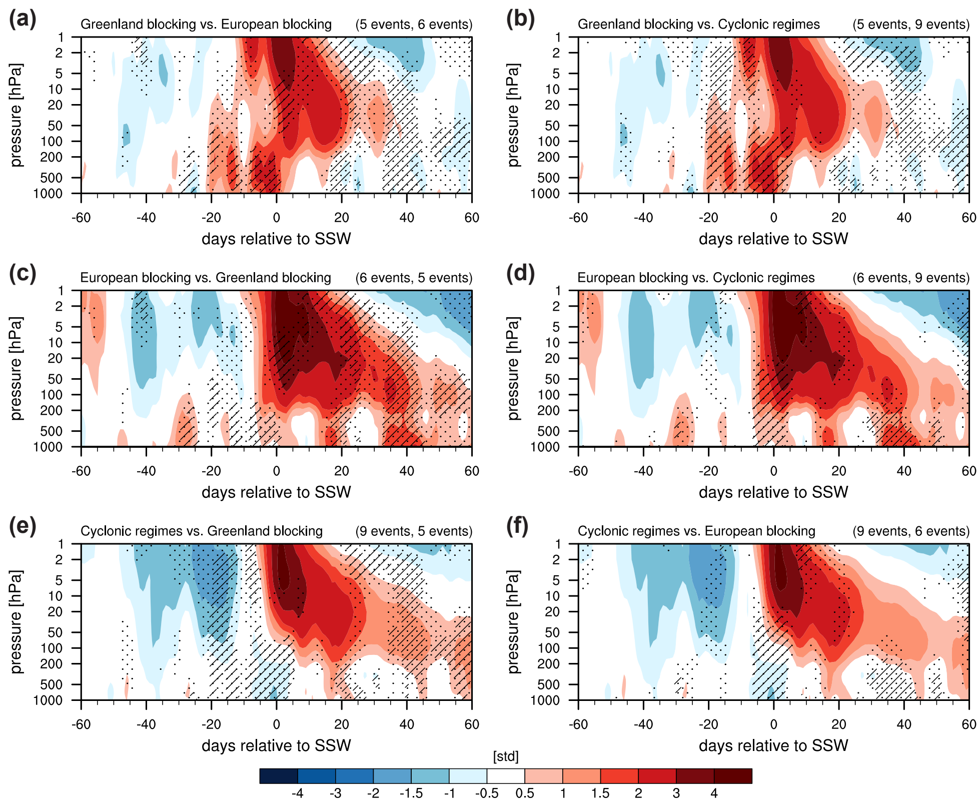

Figure 1: Standardized geopotential height anomalies as in Fig. 3 in the manuscript for (a, b) Greenland blocking cases, (c, d) European blocking cases, and (e, f) cyclonic regimes. Stippling and hatching indicates statistically significantly different anomalies from the other two samples (as indicated in the title of each panel) obtained from an overlap of the confidence intervals by less than 25 % and 10 %, respectively.

The reviewer suggests to test for the null hypothesis that the geopotential height anomalies between samples are identical. This null hypothesis may be appropriate if we were interested in the question whether geopotential height anomalies between the samples are different from each other. While this may be an interesting alternate path of investigation, it is not the question we pose here. For the reviewer's interest, we have nevertheless done mutual tests of the Greenland blocking, European Blocking, and the cyclonic regimes samples (Figure 1). This reveals that the geopotential height anomalies in the European blocking case, for example, are significantly different from the Greenland blocking case. The difference is less significant with respect to the cyclonic regimes samples (especially the anomalies before day 20).

We would like to stress that mutually testing the individual samples for difference does not a priori tell anything about whether the flow evolution in the presence of an SSW is different from that in the absence of an SSW. Similar significant / insignificant differences between samples may in principle also arise from samples obtained from random days with a given weather regime irrespective of whether an SSW occurred or not.

We hope that with these additional clarifications, we are able to convince the reviewer that our testing procedure is appropriate for the questions addressed in the study. In the revised manuscript, we will more carefully explain the statistical testing procedure in subsection 2.2 and discuss the appropriate null hypothesis.

Further, we do not agree that the no regime category would provide a suitable null hypothesis for the hypotheses addressed in this study. First, the no regime category arises due to a weak projection of the geopotential height anomalies in the Atlantic European sector on one particular cluster centroid. This does not necessarily imply that the geopotential height signal is particularly weak. Instead, this may just reflect a progression between regimes. Second, we aim to test whether the regime succession is different in the presence of an SSW compared to the case where no SSW took place. Thus, the null hypothesis is that the regime progression in the aftermath of an SSW is the same as in cases with the same initial regime but no SSW.

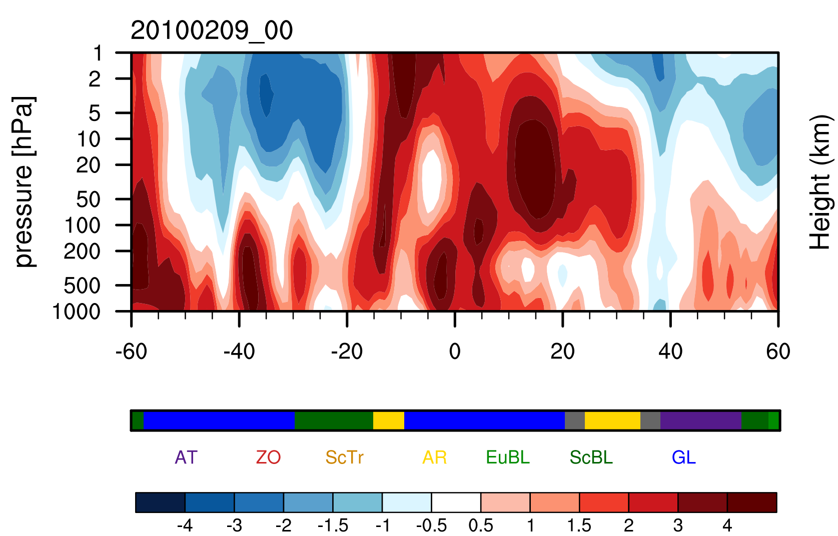

Figure 2: As Figure 4 in the manuscript but for the SSW on Feb 9, 2010.

4. **Case studies:** As suggested by the reviewer we have here included the dripping paint plot for the 2010 SSW event (Figure 2). The event again supports the general structure of the events that

are dominated by GL at the onset of the SSW, with a strong tropospheric signal already before the onset, and the tropospheric response generally limited to the weeks just after the SSW event and including an AT response. We agree that the event exemplifies variability between events in the same sub-category, but this event as well as all others in the GL category indeed show a very different response to e.g. the EuBl cases. We will be more careful in presenting the case studies (if any) and reformulate this section.

5. **Persistence of the tropospheric weather regimes:** We will clarify the relevant timescales and persistence of the regimes in the new version of the manuscript. In particular, we are indeed not suggesting that the response 6-8 weeks after the SSW onset is dominated by the weather regime at lag 0. We will formulate this more carefully in the revised manuscript.

6. **Use of pre-satellite date:** Thank you for this comment. We had considered including pre-satellite data for the original manuscript, but we were worried that the atmospheric state would be much less constrained and rely more on the model itself, hence we decided against using additional SSWs with a poorer representation of the atmospheric circulation.

7. **References / technical comments:** We will of course be happy to discuss the additional references suggested by the reviewers in our revised manuscript, as well as the minor and technical comments by both reviewers.