Response to Reviewer 1 Comments

Overall comments: *This is an excellent, well-written paper that contributes to the observation and diagnostics of Sudden Stratospheric Warmings (SSW). I have no substantive comments, and recommend that the paper be accepted after consideration of my editorial comments, which I have made on the attached file. Most of these are suggested wording changes for consideration by the authors.*

There is no need for me to review the revised version.

Response: We thank Reviewer 1 for the overall very positive commenting, which we really appreciate. We have also carefully accounted for the editorial comments as suggested by the reviewer. Please see the updates at the corresponding text locations in our Revised Manuscript (RM).

Response to Reviewer 2 Comments

Overall comments: *Li et al. use stratospheric temperatures provided from GNSS-RO measurements and ERA5 reanalysis data to propose a new definition for sudden stratospheric warmings (SSWs). They additionally define a series of diagnostics which can be used to further characterize the SSWs by their duration, amplitude, and areal extent. Their results show examples of how events defined using ERA5 compare to those defined using GNSS-RO, the time evolution of their diagnostics for select SSWs, and statistics of events aggregated over the 42 winters from 1980-2021.*

The proposed SSW definition and characterization diagnostics are new, but there are a number of issues with the way the authors have described their methods and characterized their results. Their manuscript requires substantial major revisions before being acceptable to publish. Below I provide my questions, comments, and concerns on the manuscript:

Overall Response: We thank Reviewer 2 for the very valuable comments, which are highly appreciated for their triggering of a range of further improvements to the manuscript, including the aspect of flexibility regarding the vertical resolution of input data fields that we original intended to leave for follow-on study. We have carefully addressed all comments as stated below and as reflected in the corresponding updates in the Revised Manuscript (RM). The track-changes version of the RM explicitly shows all the updates implemented.

General comments: (GC1) *In my opinion, it's not really clear what the manuscript has to do with climate change beyond showing some results in a single figure that show decadal trends. Reading the title of the paper alone would lead me to believe there should be more analysis related to climate change, when in fact the paper is primarily focused on demonstrating a new SSW definition. Personally, I think the paper should use a different title that is more appropriate for the content it provides. For example "Characteristics of sudden stratospheric warmings defined using reanalysis and radio occultation temperature data", or similar.*

Response GC1: Thank you for this comment, which we do understand and we did consider various options to update the title, in view also of the updates implemented in the RM. Since we placed a focus in the updates to further emphasize and enhance the long-term trend results related to the recent climate change period since 1980 we finally chose to only slightly change the title in form of inserting “(1980-2021)”. We think, this adds a bit of more touch of “just” monitoring (and of course inspecting characteristics, etc.) over 1980 to 2021, in addition to our intention to emphasize that we indeed see initial evidence for actual climate change dependence. To this latter end, we have split out in the results description a dedicated new Section 5 on “SSW climatology and trends under climate change” from the previous Section 4, and expanded the trend analysis so

that in the RM a separate dedicated Figure 11 is included and discussed (expanded from the previous Figure 9e). This at the same time also responded to aspects suggested by another reviewer related to the original Section 4. For the specific updates please see the Sections 4 and 5 in the RM and also our further responses below, in particular the next one to GC2.

(GC2) *This paper and prior papers have listed valid criticisms of different SSW definitions. However, the new definition proposed here sticks out as being extraordinarily. For one thing, it's not easily apparent what the authors even propose is their singular definition of an SSW. I would expect this kind of information to be mentioned clearly in the abstract since it's fundamental to the subject of the paper and for others to use the definition. Instead, one has to work backward from all the information in the long and complicated tables 1 and 2. Specifically, in table 2 item 4 we see the SSW detection criterion is based on SSW-MPD >= 7 days. What is MPD? It's main-phase duration (table 2, item 1), which is based on the number of days with SSW-MP-TEA available. What is SSW-MP-TEA? It's the main-phase threshold exceedance area (table 1, item 9), which is the maximum of the PP-TEA and SP-TEA. SP-TEA (table 1 item 8) then depends on LSTA-TEA (table 1, item 3), and PP-TEA depends on MSTA-TEA (table 1, item 2), which each depend on different temperature anomaly thresholds depending on the level of the stratosphere. There's so many steps and acronyms here that makes it difficult to even know what the SSW definition is when you've reached the end. The CP07 SSW definition involves somewhat subtle details (which prevent double-counting and exclude final warmings), but I can at least walk away with the simple "reversal of 10 hPa, 60°N zonal mean zonal winds".*

It's also not clear whether the proposed definition lends itself to being applied to other datasets. What vertical resolution of data is required for this definition? Climate model and subseasonal-to-seasonal forecast datasets are commonly output on a very limited set of pressure levels -- can this method still be used with, e.g., only 50 and 10 hPa levels? If the definition can only be readily applied to reanalysis data for monitoring purposes, then that really limits its utility.

Response GC2: Thank you for these quite substantial comments. At first glance we thought this is “too much”, especially the issue of which resolution (or how many levels) temperature input data sets would need to supply, which are aspects we had intended to leave out form this study (for the sake of limiting the work). However, we decided that your comments let trigger us that we expand our internal testings that we had in this direction to an additional comprehensive substudy, from which we can include the results in this paper already.

This led to the new Section 5 in particular, and to new Figures 10 and 11 (replacing the original Figure 9). It also let to text updates at various places throughout the RM (e.g., updates to Tables 1 and 2 and explaining in Section 3.2 that we use the ERA5 data as a basic case [still] at full vertical 137 level resolution, but for crosscheck also use the standard 37 pressure level products available via the C3S, or as a minimum-case a

simple two-levels approach with 10 hPa and 50 hPa levels only). Overall, these updates show that our method can be robustly used also for the coarser resolution data, and also for just two-level 10 hPa and 50 hPa data (see also second part of this response in next paragraph). Along with the additional substudy and its additional comprehensive sensitivity testing for these updates, we have also updated a few of the parameters (in Tables 1-2), in particular for users with simple two-levels input, and have chosen a recommended detection criterion MPD of 6 days rather than 7 days now (to best serve all input resolutions). Some specific details of the related sensitivity tests are also noted in our response to GC3 below. We really like to thank you for motivating us to include this aspect of "flexibilization" of our method with regard to temperature input data sets already in this paper. Please see the RM for all the respective updates (track-changes version RM explicitly showing everything).

On your second (or rather first) main point in your GC2 comment: yes, we also do agree that our proposed more elaborated method is less straightforward to apply (and apparently also to describe, despite our efforts to be clear) than more simplified methods. And we agree—not least based on your quite eloquent description of a possible reader's experience in trying to understand our method from detailed study of Tables 1-2—that a further effort to make our description more accessible is needed. In the RM, we hence included to this end updates to the methodology description (Section 3.3) and a related new Figure 2 that, as an overall introduction, provides a schematic overview of the workflow (main algorithmic steps) of the new method. This now quite facilitates for the readers to get a "one stop overview" first, before being invited to learn the details from the (updated) Tables 1-2. We note that these updates also account now for the point, that it is clear for the users that temperature input fields can either be used at high vertical resolution (basic case, later in Section 5 a recommended case if such data are available as option), some coarser resolution standard pressure level data (like the ERA5 37 pressure level data), or just 10 hPa & 50 hPa level temperature data.

(GC3) *The authors' justifications for the various thresholds they use in their SSW diagnostics are lacking. The authors note: "we here made sure for long term application that the four SSW TEA key variables are captured and exploited in a way so that they reliably detect and quantify actual SSW warming" and "Extensive robustness and sensitivity testing provided us with due evidence and confidence that these characteristics should enable a new level of quality and quantitative insight into SSWs". These statements do not really tell the reader why you specify these many thresholds in Tables 1 and 2:*

At least +30 K anomalies for MSTA-TEA
At least +20 K anomalies for LSTA-TEA
At least +40 K anomalies for USTA-TEA
At least $3x10^6$ km^2 >= 3 days for PP-TEA
At least $3x10^6$ km^2 >= 5 days for SP-TEA
At least $3x10^6$ km^2 >= 21 days for TP-TEA
At least 7 days for SSW-MPD for the SSW detection criterion

Below 90x10^6 km^{2}days for minor classification
Between 90 and 180 million km^2 days for major classification
Above 180x10^6 km^2 days for extreme classification
Less than/greater than 21 days of TPD to define a non-TC/TC event.

Response to GC3: Thanks for this comments, where we again agree that it is a number of parameters along the algorithmic steps of the method that appear here, for which a reader perhaps would like to be informed on the rationale to a level so that she/he can have due confidence that we have done properly careful work to deliver robust choices here. In the RM, we responded to this comment in that we further clarified issues in Tables 1 and 2 as well as in the related text, which is in particular the Section 3.3 on methodology. While we quite extended also the text accordingly, and also included the new Figure 2 to facilitate the basic method understanding (see previous comments above), we prefer not to overload the paper with exceeding details on the quite massive work that we put into the comprehensive sensitivity testing. Please see the RM for how we handled this (also thinking of limiting the overall length of the paper).

We have of course documented everything quite well, in internal "reporting documents" that are many pages long. We provide here a brief summary, for some more detailed infos to the reviewer also beyond what we decided to explicitly include in the RM:

The thresholds we selected in Tables 1 and 2 are determined based on large sensitivity tests. The determination of thresholds includes several steps. We first carefully analyze magnitudes of temperature anomalies and TEAs to understand variations in different SSW cases, initially working with limited ensembles of events and then with the full 1980-2021 data record, and then parameters are evaluated and selected. The basic criteria of selection is that the main warming/cooling features should be captured and our metrics can reliably detect and quantify actual SSW events. The detection results of the BG18 climatology, which we consider a very good work, are used as sort of a reference in our determination of selecting thresholds. Main sensitivity tests we made on determining thresholds and the further metric choices include:

(1) Sensitivity tests of temperature anomaly thresholds.

For chosen anomaly thresholds for the TEA key variables, we have used 25 to 35 K for SSW-PP-TEA and 15 to 25 K for SSW-SP-TEA, and found 30 K most suitable for SSW-PP-TEA and 20 K for SSW-SP-TEA, respectively (limited tests on other thresholds informed the reasonable test range). Practically, we found that in middle stratosphere layer, temperature anomalies should not be smaller than 30 K otherwise the warming is found all over the polar region and the characteristic SSW patterns" cannot be captured. Similarly, in lower stratosphere, temperature anomaly should not be lower than 20 K for the same reasons. On the other hand, if a higher threshold is chosen it does limit the sensitivity to robustly capture minor SSW events. In the upper stratosphere layer, most upper stratospheric trailing-phase cold anomalies (the purpose of this layer) are found to reach within -30 to -40 K and we hence decided to choose -30 K for SSW-TP-TEA. Overall, changing these thresholds somewhat, little impacts are found on those that we classify into major and extreme events, hence they are tuned to be adequate to get the minor events also captured down to a sensitivity

level, below which detected anomaly patterns would be doubted to be SSW features.

(2) Sensitivity tests of duration choices.

We have tested 3 to 5 days for SSW-PP-TEA and 5 to 7 days for SSW-SP-TEA. For the duration of combined SSW-MP-TEA, i.e., SSW-MPD, from 5 to 9 days (additional to the 6 days finally selected). It is found that 3, 5, and 6 days are most suitable for SSW-PP-TEA, SSW-SP-TEA and SSW-MP-TEA, respectively, which we tested over the complete record. More limited tests on smaller numbers of events showed that other duration choices out of these test ranges are rather obviously suboptimal for the SSW detection and monitoring purpose. Durations with a day less or more for the SSW-MPD (i.e., 5 days or 7 days instead of the 6 days) are clearly possible to be used and, as one might expect, don't change the results or conclusions in any drastic way, but we found them already somewhat less optimal for all the different input data resolutions (see point (4) below). Practically, if allowing a bit shorter duration, more minor events will be included compared to our existing results (starting to be doubtful whether they really qualify as "SSWs"). If using a bit longer duration, some minor events will be excluded (also a few that clearly look like "SSWs" also on detailed space-time tracking inspection). Major and extreme events (in our suggested classification, see next point (3)) are rarely influenced anyway.

(3) Sensitivity tests for SSW classification.

Classification criteria are determined based on two reasons. First of all, numbers of events determined by our classifications should not be largely sensitive to the selected thresholds, though clearly there are no hard physics-based limits strongly constraining the selected classification boundary values (though, e.g., the appearance of "trail-cooling" is a quite clear marker of the class of "extreme events"). We however find it is useful to have some few-classes distinction and that is why we suggest it. Secondly, the determined major and extreme events should be reasonably insensitive to thresholds selection for TEA and also to the vertical resolution of input data. This is why we, for the simple two-levels 10&50hPa approach, recommend to use the lower 70/140 classification boundaries rather than 90/180 (also seen in Figure 10e vs Figure 10a in the RM); see next below also point (4) on these sensitivity tests. We do find the classification quite helpful, although especially the boundary between minor and major event class is somewhat more subjective, in being chosen half-value of the upper boundary between major and extreme events. However, the results support that the event frequency found for major+extreme events is quite insensitive in this way, for example, on the vertical resolution of the input data, event though the simple two-levels approach overall detects less events (see next point (4)).

(4) Sensitivity tests on vertical resolution.

In order to make sure that our method is also applicable for data with less vertical resolution, we made further sensitivity tests. First of all, we formulate MSTA-TEA using temperature anomalies at 10 hPa instead of 30-35 km and also formulate LSTA-TEA using anomalies at 50 hPa instead of 20-25 km. Since the corresponding altitude of 10

hPa (about 31 to 32 km) and 50 hPa (near 21 km) is quite low within our original altitude layers, therefore temperature anomalies are smaller and subsequently TEAs are smaller. Therefore, we also somewhat lowered down the minimum TEA_{Min} (testing different values) and the classification boundaries (testing again different values). This is reflected in the updates of Tables 1-2, especially the footnotes. It is found that the detection results on major and extreme events are quite consistent, and most of the minor events are included. Clearly a number of (minor) events is missed against the basic case, but this is to be expected and is explained in the RM.

Additionally, regarding the vertical resolution, we have also tested the coarse ERA5 37 vertical pressure levels by formulating MSTA-TEA using linear-interpolated temperature anomalies from pressure levels and computing then the layer-mean within 30-35 km and formulating LSTA-TEA similarly for within 20-25 km, and USTA-TEA for 40-45 km. The results are found close to our ERA5 137 vertical level basic-case results. Since ERA5 37 vertical pressure level data is free to all users from the C3S, people can freely get the data online and use for our method for SSW detection. Also they may use a bit more sophisticated vertical interpolation, which brings the results even closer to the basic-case results.

(GC4) *While I can generally understand what the authors intend to say throughout the paper, the text will overall require a significant amount of editing for grammar. In my specific comments below, I have primarily focused on asking substantive questions/comments rather than pointing out grammatical issues.*
Response to GC4: Ok, we have made another effort throughout to include editorial updates to improve the grammar and the English in general. Also the other reviewers had some language improvement suggestions that were accounted for in the RM.

Specific Comments (L# refers to the line numbers of the manuscript):

Point 1: *L34: It's not clear in this sentence, but the westerly winds are the polar vortex. The westerly zonal mean zonal winds of the polar vortex can reverse during a strong (or major) SSW, but the three-dimensional polar vortex can undergo a displacement or split.*
Response to Point 1: Ok, we have improved this sentence accordingly in the RM.

Point 2: *L35-36: This sentence is very unclear. The planetary waves from the troposphere can be modulated by the QBO, ENSO, etc.*
Response to Point 2: Ok, we have improved also this sentence in the RM.

Point 3: *L77-80: I don't agree with these assessments. Plenty of studies have been done that are able to draw robust conclusions about weather and climate phenomena based on simple SSW definitions.*

Response to Point 3: Ok, we have updated the text here and have somewhat "toned down" our assessment, better appreciating the (undoubted) value of existing studies.

Point 4: *L84: Earlier on you make the argument that reanalyses have inhomogeneieties and irregularities due to observing system changes; apparently this is not a big enough deal to prevent the usage of ERA5 for purposes of defining/characterizing SSWs?*

Response to Point 4: Yes, this is true, since it appears that the SSWs are "strong anomaly" features, while the residual long-term inhomogeneities in (ERA5) reanalysis temperature fields are quite small in size. Hence these inhomogenieties somewhat effect classical long-term "mean field" temperature trends but are small relative to such "strong anomalies". We have updated the related textpiece to point to the value of verification of the reanalysis data by long-term-stable observational (RO) data.

Point 5: *L181: How do previous climatologies lack quality during the 1990s? Figure 1: How are you determining the number of profiles for ERA5? With a resolution of 2.5x2.5 degrees, you have 144 lons and 73 latitudes, which corresponds to (144*71) + 2 = 10226 profiles over the entire globe (the poles only count as single points).*

Response to Point 5: Ok, we have improved the sentence with the "previous climatologies" to a more considerate formulation. On the number of profiles of ERA5, we have a profile at every 2.5 °×2.5 °grid point within 50 °-90 °N (16 x 144 grid points) and we have it at four analysis times per day, leading to the "near 10000 per day" (actually it is 16 x 144 x 4 = 9216). Similarly, over 60 °-90 °N you arrive at 6912 profiles, which is why the pink line in Figure 1 appears close to 7000. We have inserted a hint in the caption to say it's "four analysis times per grid point per day".

Point 6: *L248-255: This description of how you made decisions is not really sufficient. How were the thresholds in table 1 and table 2 chosen? These details should be available to the reader within this paper, since you are proposing your definitions be standard.*

Response to Point 6: Ok, as discussed in the responses to the general comments above (see responses to GC2 and GC3) we have substantially updated this description plus the Tables 1-2 plus inserting a new Figure 2.

Point 7: *L265-269: Doesn't ERA5 assimilate the RO measurements? If so, then it is not really surprising that they agree well.*

Response to Point 7: Yes, and we mention and discuss this. The one dataset just verifies the other (in how they come up with the SSW patterns, etc.); it is not a validation between independent datasets. The verification helps establish confidence that for such analysis of "strong anomaly" patterns, the ERA5 is good enough and can do a decent "job" also for long-term, even if it has some (comparatively small) inhomogenieties.

Point 8: *L283-287: Why do your diagnostics not characterize the depth of significant*

warming? SSWs that lead to persistent temperature anomalies in the lower stratosphere are thought to be the ones most likely to lead to coupling with the troposphere.

Response to Point 8: We consider that we do characterize the "depth" to some degree by our discussion of Figures 4 to 9 (new figure numbers in the RM) in Section 4, since we have characterization metrics also on "depth" and show some results for TEAs at different threshold levels. In fact we did check for long-term changes also of "depth information", using our auxiliary metric on "SSW onset maximum warming anomaly" (plus our background data providing a daily tracking of this max. warming anomaly during events), but it did not really add much to the TEA info (i.e., was found quite correlated, according to a "the larger the deeper" rule; we mention as part of the results). Having said this, we confirm that a closer investigation of "depth", or of "exceedance volume" as we also explore, is quite interesting nevertheless. It is therefore part of our follow-on work, since it is a significant extra effort beyond the scope of this study.

Point 9: *L291-292: This is by definition of how you characterize minor, major, and extreme events.*

Response to Point 9: Ok, yes, and we now point to this by an insert in this sentence.

Point 10: *L349-351: Your statistics count all of your SSWs, including minor events. The numbers you are comparing are not directly comparable since, e.g., the 0.6 event/year figure corresponds to a particular definition for *major* SSWs. It would be more useful and more interesting if you listed the individual frequencies for your classification of minor, major, and extreme events (or major+extreme).*

Response to Point 10: Thanks for this good suggestion. Ok, we now include in the Figure 9 update (now the Figure 10) an extra frequency estimate also for major+extreme events and discuss it accordingly in the results section (now in Section 5). This additional frequency has helped also our description related to the different vertical resolutions of the ERA5 input data fields that we now intercompare (Figs 10-11).

Point 11: *L365: I do not think this is correct. The 90s were a notably cold/quiet period for winters in the NH stratosphere. While minor-warming-like events did occur, there were not many that were associated with zonal wind reversals. This shouldn't have much, if anything, to do with the number of radiosonde stations.*

Response to Point 11: Ok, thank you, we have reformulated this paragraph to a more considerate formulation in the RM, which notes several aspects that may contribute to the difference.

Point 12: *Table 3 and Figure 9: Is it really appropriate to "double-count" with events that occur within 1-2 weeks of others? Most definitions of SSWs take into account that radiative timescales are pretty long in the stratosphere such that potential events that occur within ~3 weeks of another are considered to be part of the same event. By my reckoning of Table 3, there are 4 events which occur very closely in time to others (e.g.,*

1989-02-12 and 1989-02-20, 2000-12-07 and 2000-12-18, 2003-12-24 and 2004-01-04, 2006-01-11 and 2006-01-21). How are the results of Figure 9 impacted if these are not "double counted"?

Response to Point 12: Thanks for these good comments; we had been considering and inspecting this type of aspects in great detail as part of our comprehensive testing work, also including detailed inspection of event sequences within a winter. We understand that, for example, most wind reversal-based definitions will not include events separately that occur within ~3 weeks of another. However, based on our inspections as part of developing the method we propose, we prefer to keep those detected by our method as separate (minor) events, since inspections suggest them to be individual (minor) SSWs with their own primary phase, etc.

As one example for the reviewer, Figure R1 shows a sequence of temperature anomaly contours of MSTA from 2000-12-05 to 2000-12-10 and Figure R2 shows temperature anomaly contours of MSTA from 2000-12-16 to 2000-12-21. These two temporally adjacent warmings occurred with locations centers at very different geographic regions. Hence one might prefer (as we do) to say that these two temporally close events are considered separate events. We think that one of the main basic purposes of SSW monitoring is to characterize polar winter stratospheric variability, which we also see better characterized if keeping (minor) events (as found by our temperature field-based method) recorded as separate events.

Regarding the impact on Figure 9 and specifically the long-term trends, our robust form of computation of the decadal-mean values (see first paragraph of now Section 5.2 in the RM) does not depend on the event count and hence these results are almost unaffected by our choice of counting these events separately or not. Regarding simply the numbers counted, the minor event number would be obviously impacted, while the event frequency of the major+extreme events only would be almost unaffected.

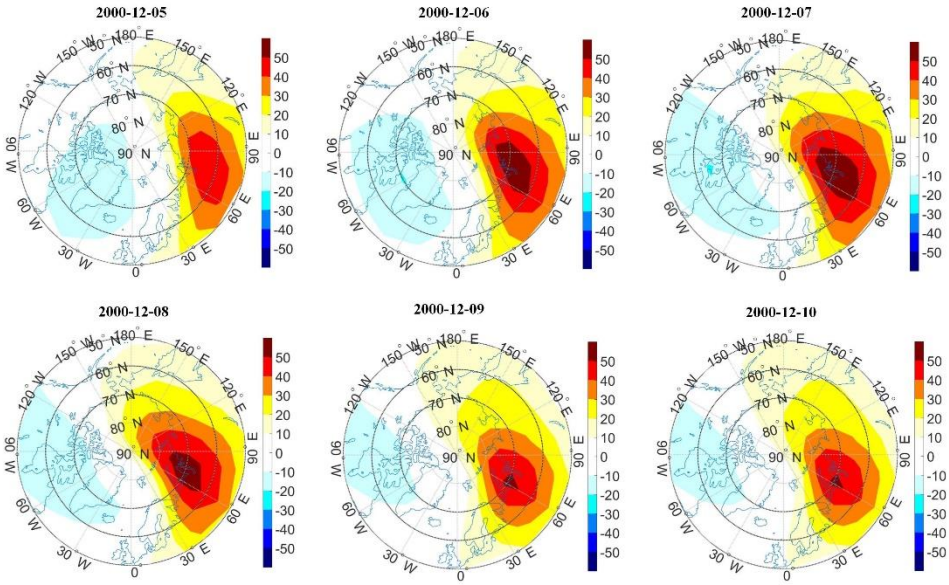

Figure R1. Temperature anomaly contours of MSTA from 2000-12-05 to 2000-12-10.

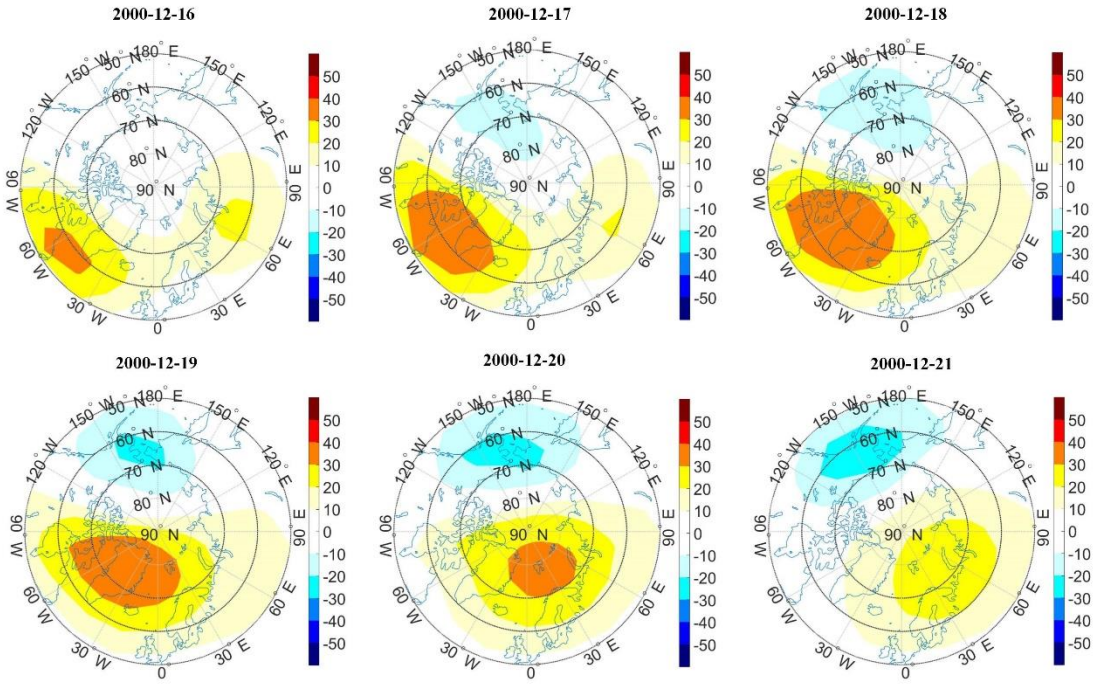

Figure R2. Temperature anomaly contours of MSTA from 2000-12-16 to 2000-12-21.

Point 13: *Figure 9e: I am confused at what is plotted in these panels. Each of the points are decadal averages? How do you account for gap years (i.e., years in which you do not detect anything) and years with more than 1 event? You always divide by 10 years, rather than the number of events in the given decade? How sensitive are these results to the fact that you require a MPD of 7 days for an event to even be considered? I.e., how much do your results change if you instead considered events with at least a MPD of 5 and 9 days?*

Response to Point 13: Ok, we understand that the previous Figure 9, including Figure 9e, was quite a bit overloaded and the related text in previous Section 4 was not informative enough. As already mentioned in our responses to the general comments above, this part of the paper hence was the most strongly updated one, including also based on comments by another reviewer. In particular, the information of previous Figure 9e was dropped from Figure 9, which became a new Figure 10 not including the long-term trend results. The long-term trend results received a separate dedicated Figure 11 (showing now the results for different input vertical-resolution options), and also the whole description is now placed in a separate dedicated Section 5. Please see the RM (track-changes version) for the details of all these updates.

Regarding the sensitivity to the choice of the MPD, please see our response to general comment GC3 above. As noted and explained, we now recommend resp. adopted as the overall choice that best fits all vertical-resolution input fields an MPD of greater or equal 6 days as the SSW detection criterion (previously we had chosen 7 days, where we only had discussed the full-resolution basic case). On our comprehensive sensitivity studies backing all this please also see our response to comment GC3 above.

Response to Reviewer 3 Comments

Overall comments:

This paper uses a new approach to monitoring the SSW and uses this new approach to understanding the climatology of SSW. This paper provides a useful tool for future SSW studies. The science of this paper is interesting, and most of the text in this paper is well structured, for example, the introduction and the conclusions. However, the analysis in this paper is very unclear and many places hard to read, and I recommend a major revision before accepting. Most of my comments are only related to how to clarify it and not about the science, so I believe the authors will eventually make it a publishable paper.

Overall response: We thank Reviewer 3 for the valuable comments. We have carefully addressed all comments as stated below and as reflected in the corresponding updates in the Revised Manuscript (RM). The track-changes version of the RM explicitly shows all updates implemented.

General comments:

Point 1: *Many figures in this paper are (a) too complex, and (b) of poor quality. For example, in figure1, 5, 6, and 9, the annotations overlap with each other so very hard to recognize. Especially in figure 1b-d, with so many large dots, the readers cannot read any information. In figs. 6&7, figure y-axis limits are too low, and some data is cut off. The authors should really find out a way to convey the information in your figures clearly and explicitly. At least, your text in the figures should be easy to recognize.*

Response 1: Thank you, ok, we have basically put a lot of care into the design and making of the figures and have now made a further significant effort to improve in this direction. We added a schematic overview figure on the method (related to a comment of another reviewer), replotted figures 1, 6, 7 and 8, and decomposed Figure 9 into two figures (now Figures 10 and 11) in the RM, to avoid any annotation overlaps and to make the information clearer to readers, including more recognizable texts.

Point 2: *Many paragraphs in section 4 should be rewritten. I recommend using an opening sentence to state the argument of this paragraph, instead of saying 'Figure 3 shows…'. What is important in your paper should be these scientific arguments, not the explanations of your figures. In my opinion, sections 4.1-4.3 are only listed results, and*

section 4.4 should be the scientific conclusions you should emphasize, so efforts are needed to re-organize the paper and extract out useful information.

Response 2: Thanks, we made also a significant improve effort in this direction (though we respectfully disagree somewhat that 4.1-4.3 are "only listed results"). We further improved texts in 4.1-4.3 and, in particular, we also split out from Section 4 a dedicated Section 5, to highlight the climatology and long-term monitoring key results. In this latter section (where we also expanded the results themselves in related figures, based on comments by another reviewer) we as well increased the weight of scientific arguments and interpretation. Sections 4 and 5 in the RM with track changes show these updates.

Specific comments:

Point 1*: Line 90 – why there is a () in reanalysis data?*

Response 1: The () is to indicate that both reanalysis data, and (standard operational) analysis data have been used for the validation of RO data.

Point 2: *Line 150 – name-coining: what does this mean? Also, you need to rewrite this sentence, for example, you should not use (i.e.) after as such*

Response 2: "name-coining" meant that using "sudden stratospheric **warming**" as a term recognizes that it is a temperature anomaly. However, as seen in RM, we dropped the unclear "name-coning" from the sentence now and also rewrote it to be clearer (including that there is no expression with "i.e." in parentheses any more).

Point 3: *Line 152 – secondly, second after what?*
Response 3: Thanks, we have somewhat rephrased also these follow-on sentences, so that "secondly", and "thirdly", do not exist any longer in the RM.

Point 4: *Line 180 – 'previous published climatologies reach to 2013 only and lack quality over the 1990s decade' I think it is not true. Also, it should be 'previously published.*
Response 4: Thank you, we have revised also this sentence accordingly.

Point 5: *Line 241, line 265– 'are overall similar', 'appear rather similar': conclusions like 'similar' and 'appear' are too subjective and should not be in a scientific journal article, please check the rest of the paper to clarify your statements.*
Response 5: We agree that vague interpretations should be avoided as possible and we hence rechecked and changed accordingly at a number of places throughout the text.

Point 6: *Line 267 – 'leading to somewhat': delete somewhat*
Response 6: Ok, we can agree and deleted "somewhat" at this place.

Point 7: *Line 296 – 'same three events': how to define 'same'?*
Response 7: Thanks, we avoided to use the phrase "same three events" in this sentence and actually reformulated the whole sentence into a simpler and clearer statement.

Point 8: *Line 339 – this long sentence is too hard to understand*
Response 8: Ok, we agree and have now split the sentence into two sentences in the RM, and reformulated a bit, which clearly eases the understanding of this information.

Point 9: *Line 426 – 'we detected a number of events', how many?*
Response 9: Thanks, we included the specific number now in the RM (we detected seven events).