Author’s Response File, Comprising:

(1) Referee’s comments, with authors responses interspersed.

- Near the beginning of the responses to each reviewer, there is a description of the two major changes (omission of MERRA from the reanalyses compared and using an REM rather that a single reanalysis for the reference for difference plots) made in response to the reviewer’s comments.

(2) Tracked changes (latexdiff) version of the manuscript.

- Note that the tracked changes version shows only the figures for the revised version.
- Because of the extent of the changes, the tracked changes version is not always terribly helpful.

Responses to Reviewer 1's (listed as "Referee #2" on discussion site) comments

The reviewer's comments are given in *black italics* and our responses in blue plain text.

This manuscript provides an extensive intercomparison of diagnostics relevant for polar stratospheric ozone processing in five recent 'full-input' reanalyses, MERRA, MERRA2, CFSR, ERA Interim, and JRA55, as part of the S-RIP intercomparison project. The study is thorough, well thought out and generally clearly presented, and the intercomparison should provide a valuable reference point for studies of polar processing that are based on reanalysis data, as well as a reference point for comparisons of these quantities in future reanalyses. To me the more interesting results are the almost ubiquitous improvement seen in the agreement between reanalyses following the advent of improved satellite observations around 1998-2000, as well as the increased sensitivity to threshold definitions seen in the NH relative to the SH. The results are not earth shattering, but are of value and as such I would recommend publication after some minor revisions.

My main concern is that the paper is very long, and that its impact would be greater if it were significantly shorter. As a potential reference for future studies, there is some value in being rather complete in the intercomparisons, but 21 figures is a lot more than most readers will want to go through. It's not clear to me that Figs. 1-3 are really necessary, nor what is the additional gain from including Figs 18-19 over the content of Figs. 16-17.

We thank the reviewer for their helpful comments. The paper has been extensively revised in response to major comments by the other reviewer, Simon Chabrillat, so there is not a one-to-one correspondence with all of the specific suggestions made by the reviewer. We have, however, tried to keep specific new material as concise as possible and have removed material where it was suggested by either reviewer, as detailed in the specific responses below. This includes removing the original Figures 2, 3, 16, and 17 and the associated discussion.

Two major changes to the paper motivated by Simon Chabrillat's comments are to use a reanalysis ensemble mean (REM) as a reference for the comparisons rather than using MERRA-2 (see our response to Simon for discussion of this), and to remove MERRA from the reanalyses evaluated in this paper. There are numerous reasons for removing the MERRA comparisons, including the following: The choices that were made by GMAO of which products to archive for MERRA have made "fair" comparisons difficult to impossible for many products, including potential vorticity (PV), which is critical for stratospheric vortex and many other studies. While comparing MERRA with MERRA-2 and other reanalyses was critical to evaluating MERRA-2, numerous such studies have now been done; MERRA-2 was intended to supercede MERRA and sufficient evaluation of it has been done now to warrant this. Finally, especially

when using the REM as a reference, it is somewhat problematic to include two reanalyses based on nearly the same model in a comparison of just five reanalyses.
Because these two major changes, especially the switch to using the REM, necessitated a nearly complete rewrite of large portions of the text in the results section (though the final results changed very little), several of the reviewers' comments now refer to text that has been replaced, and it is not possible to document every change in detail.

Specific comments
p2 l1 There is a spurious 'data' here.

Fixed.

p2 l32: 'Best' is highly debatable here. They are a good tool, certainly, but they are not appropriate for all tasks.

We have changed this to "among the best".

p7 l6: The role of radiosondes should not be understated here – although it is not considered here, JRA55C, which assimilates only 'conventional' (non-satellite-based) observations does a remarkably good job of capturing much of the details of NH stratospheric variability.

We have added a sentence noting the importance of radiosonde inputs in the lower stratosphere, but also noting the caveat that the sonde data are sparse in the NH polar regions and very sparse in the SH polar regions.

p10 l4-19: The choice of a 5 day geometric mean here needs to be justified here. The key question is the decorrelation timescale of fluctuations in the differences between reanalyses. these could arise from a variety of processes with rather different timescales so it's not at all obvious to me what timescale is appropriate, but given that fluctuations in the physical quantities themselves (temperatures, PV) can have decorrelation timescales of far greater than 5 days this choice could be rendering the derived CIs rather meaningless. This can be checked directly by looking at the autocorrelation functions of some sample quantities.

There is also a question of just what it means for two reanalyses to be 'statistically' indistinguishable. There is an important distinction to be drawn as to whether a difference seen between two temporal averages is indicative of a systematic, steady bias between the two systems as opposed to a result of the residual over temporal fluctuations. But given that these systems are meant to capture the same atmospheric fluctuations, time-dependent differences between reanalyses are still meaningful and potentially quite relevant to know about. Just because this measure indicates that the fluctuations are of larger amplitude than the mean bias (in some statistically meaningful sense) doesn't mean the reanalysis products are indistinguishable.

We have added justification for our choice of the expected block length for the stationary resampling procedure.

Since we moved to using a reanalysis ensemble mean (REM) based on Simon Chabrillat's review, we examined the autocorrelation functions (ACFs) for the differences of the reanalyses from the REM. What we found is that the decorrelation timescales can vary and depend highly on the reanalysis, the diagnostic, the year, and the vertical level; in some cases the decorrelation timescales reach zero in a few days, while in other cases they remain well above zero beyond 10 days. As examples of this, we have attached two figures of the type we used to evaluate these timescales at the end of our responses to reviewer 1. They are large and unwieldy figures, but they show (1) the ACF of the raw diagnostics for the REM (top panel) and the comparison reanalysis (second panel; in these cases MERRA-2), (2) the ACF of the comparison reanalysis minus REM (third panel), and (3) 18 ACFs of 18 different stationary resampled (with expected block length of 5 days) difference time series. The two examples we show here are for SH maximum PV gradients for the same level (490 K) separated by just one year. You can see that for 2015, the autocorrelation of the difference time series (3rd panel) stays fairly large out well beyond 10 days; in contrast, the ACF of the difference time series for 2014 drops much faster. You can also see that even though the decorrelation time scale is quite long for 2015 and the average block length of the resampled time series is 5 days, there are still a handful of resampled cases that also have relatively long decorrelation timescales (see e.g., n = 3, 9, 12, 16, 17, and 18) -- and there are also many resampled time series for 2014 that match the much shorter decorrelation time-scale pretty well too. This is one of the benefits of using the stationary resampling procedure rather than block resampling; using random block sizes can help to create artificial time series that better match the autocorrelation "structure" of the original time series.

After making and examining these sorts of plots, we repeated our bootstrapping procedure and tested using different expected block lengths between and including 5 and 15 days. What we found is that in all cases, the results we obtained were virtually identical. Ultimately, for the results now shown in the manuscript, we increased the expected block length to 10 days since it seemed to be the most "happy medium" among the many ACFs we examined; we also doubled the number of resamples for our bootstrap distributions to 2x10^5.

Regarding your second point, we agree that our results from the bootstrapping analysis should not be used to judge the (in)distinguishability of the reanalyses, but should be limited to the "classical" interpretation of statistical hypothesis testing. The presence of an "x" on our pixel plots (null hypothesis can't be rejected) does not mean that the time series of the certain diagnostic, year, and level are indistinguishable, just that we cannot reject that the winter means are equal. Conversely, the absence of an "x" on our pixel plots (null hypothesis rejected) does not mean that there are overwhelming or large biases, just that the winter means are unlikely to be equal. The significance testing here primarily supplements the winter mean differences and standard deviations -- for example, there are many cases of the diagnostic mean differences being very small but “significant” (no “x”) alongside standard deviations that are very small, which just says that although such differences are generally small, they are persistent enough during the season such that many resamples of the time series shared that persistent (but small) difference. There are also some cases where the diagnostic mean differences are noticeably nonzero but “insignificant” (“x” is present) alongside larger standard deviations, which indicates that the variability is large enough such that many resamples do not share the structures that give rise to the real mean difference.

We have modified and double checked our text to ensure we have not included any misleading language regarding the interpretation of the statistics.

p10 l22-24: Are these averages and standard deviations taken over time (from the 12Z snapshots) within the year? Or are they taken over spatial degrees of freedom? Is the data synthetic? If not, what is actually shown?

A more general thought on this section - while I appreciate the effort to make the plots clear I wonder if it would be more efficient to simply explain this plot in the first case rather than present an example; the paper is quite long and omitting Figures 2 and 3 would go some ways towards shortening it without omitting relevant details.

The data here were synthetic and meant to represent averages and standard deviations taken over time as in the other results we show, but we have taken your advice to shorten the paper and have ultimately taken out (what were formerly) Figures 2 and 3.

Fig. 4: What is the relevance of the black lines 70 hPa and 30 hPa?

These are the selected levels for which separate line plots are shown. This is now clarified in the caption.

Fig. 5: Four digits of precision are not needed on the pressure axis labels

The labels are now limited to a single digit after the decimal point in all the figures.

p13 l3: Earlier in the text A_PSC has been used - this to my mind is more standard than A_NAT. Was the switch intentional?

We use the subscripts _NAT and _ice to convey the particular type of PSC threshold we are looking at. A note to this effect has been added in discussion of PSC thresholds in the methods section.

p15 l34: Up to 600K or so there is a significant improvement in the agreement between MERRA and MERRA 2 (in means and standard deviations) after 2000 - it’s just in the upper stratosphere (particularly 660 and 700K) that the disagreement becomes if *anything larger.*

The MERRA comparisons have been removed from the paper for the reasons stated in our response to Simon Chabrillat, so this text has been removed.

p16 l2: Is this a result of a more or less constant PV offset across the polar regions or differences in the locations of the maximum gradient?

This text has been revised to reflect the individual calculations of the vortex edge location for each reanalysis. The results now suggest that this is related to differences in the locations of the maximum PV gradients, which is noted in the revised text.

p16 l23: 'Total days' is a strange unit here since it's regularly far in excess of the total number of days in a year. The appropriate unit should be pressure-level days, I suppose.

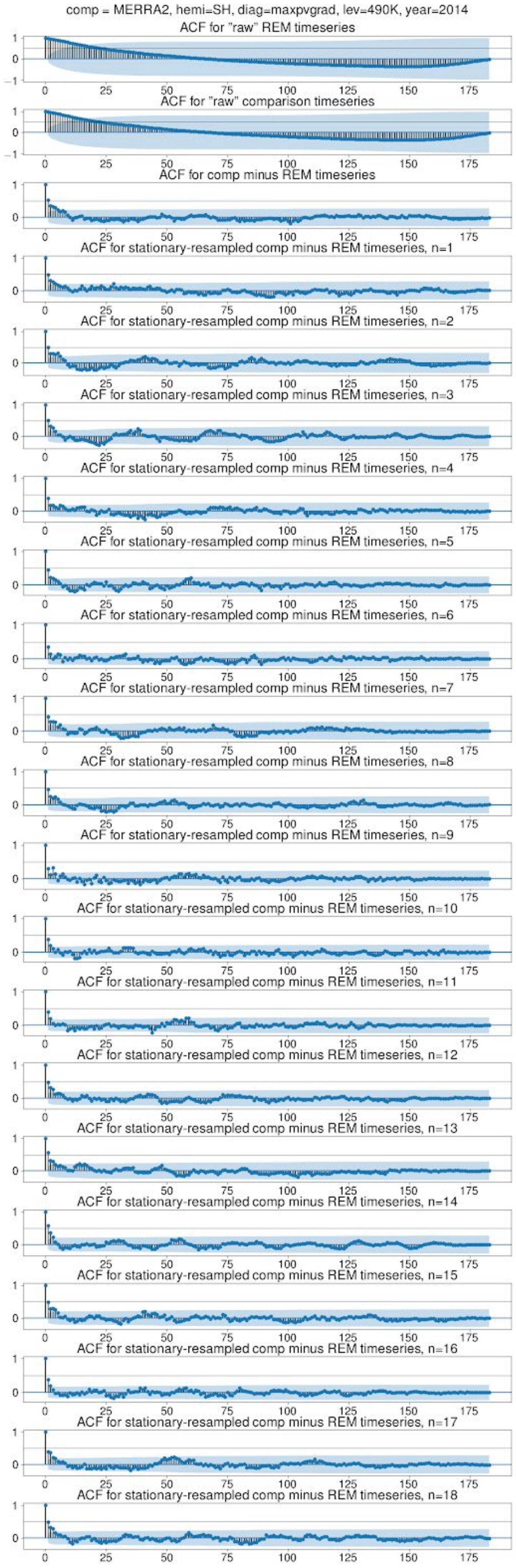
Because the V_PSC / V_vort figures provide much of the same information, and to shorten the paper, we have followed your suggestion to delete the plots showing days integrated over the levels, so these figures have been removed.

p18 l26: I can't find an explicit definition of A_vort, though there are some relevant details in section 2.2.2

Since we do not use "A_vort" elsewhere in the paper, we now simply refer to it as "vortex area". We have also made the definition of vortex area more explicit. However, please note that the paragraphs discussing the methods behind the derived diagnostics have been moved to a new subsubsection of section 2 (in response to a comment by Simon Chabrillat).

p22 l34: Given the statement two lines earlier about the similar timing of changes in the observations being assimilated by different reanalyses, the consistency of trends across multiple reanalyses should not be seen as any kind of definitive indication of the reliability of trends.

Agreement across reanalyses would be a ***necessary*** condition to believe trends derived from them to be reliable. We agree that it is certainly not a ***sufficient*** condition. We have reworded the sentence in question to make this more explicit.



comp = MERRA2, hemi=SH, diag=maxpvgrad, lev=490K, year=2014

ACF for "raw" REM timeseries

ACF for "raw" comparison timeseries

ACF for comp minus REM timeseries

ACF for stationary-resampled comp minus REM timeseries, n=1

ACF for stationary-resampled comp minus REM timeseries, n=2

ACF for stationary-resampled comp minus REM timeseries, n=3

ACF for stationary-resampled comp minus REM timeseries, n=4

ACF for stationary-resampled comp minus REM timeseries, n=5

ACF for stationary-resampled comp minus REM timeseries, n=6

ACF for stationary-resampled comp minus REM timeseries, n=7

ACF for stationary-resampled comp minus REM timeseries, n=8

ACF for stationary-resampled comp minus REM timeseries, n=9

ACF for stationary-resampled comp minus REM timeseries, n=10

ACF for stationary-resampled comp minus REM timeseries, n=11

ACF for stationary-resampled comp minus REM timeseries, n=12

ACF for stationary-resampled comp minus REM timeseries, n=13

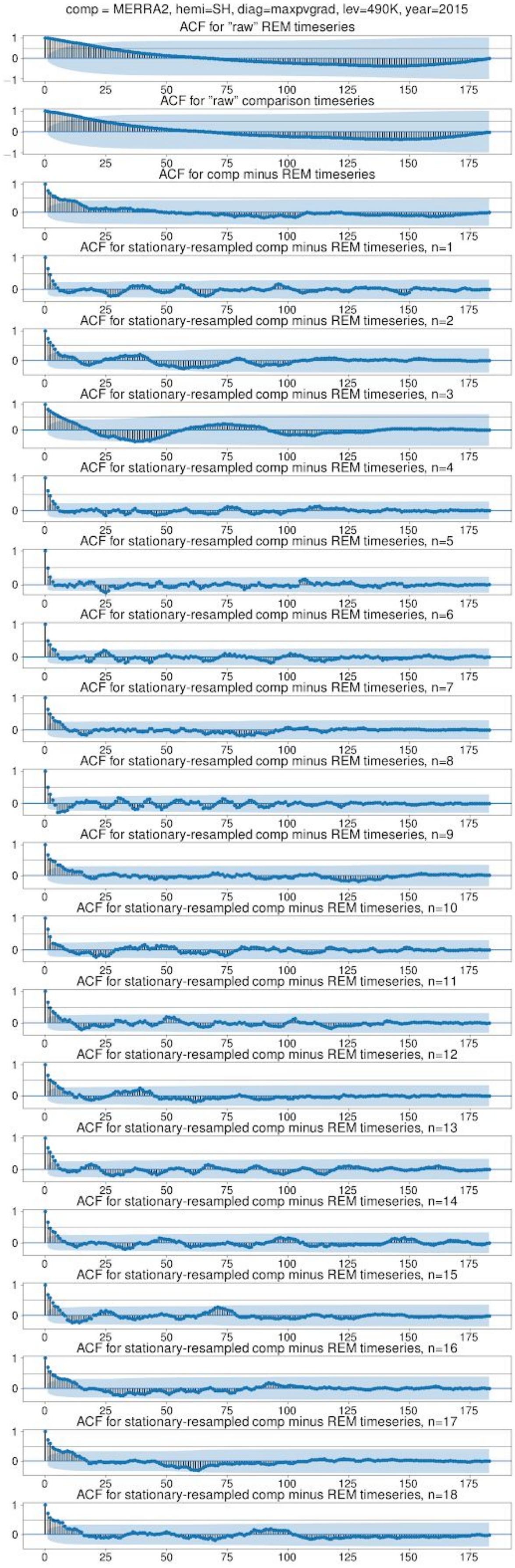
ACF for stationary-resampled comp minus REM timeseries, n=14

ACF for stationary-resampled comp minus REM timeseries, n=15

ACF for stationary-resampled comp minus REM timeseries, n=16

ACF for stationary-resampled comp minus REM timeseries, n=17

ACF for stationary-resampled comp minus REM timeseries, n=18



comp = MERRA2, hemi=SH, diag=maxpvgrad, lev=490K, year=2015

ACF for "raw" REM timeseries

ACF for "raw" comparison timeseries

ACF for comp minus REM timeseries

ACF for stationary-resampled comp minus REM timeseries, n=1

ACF for stationary-resampled comp minus REM timeseries, n=2

ACF for stationary-resampled comp minus REM timeseries, n=3

ACF for stationary-resampled comp minus REM timeseries, n=4

ACF for stationary-resampled comp minus REM timeseries, n=5

ACF for stationary-resampled comp minus REM timeseries, n=6

ACF for stationary-resampled comp minus REM timeseries, n=7

ACF for stationary-resampled comp minus REM timeseries, n=8

ACF for stationary-resampled comp minus REM timeseries, n=9

ACF for stationary-resampled comp minus REM timeseries, n=10

ACF for stationary-resampled comp minus REM timeseries, n=11

ACF for stationary-resampled comp minus REM timeseries, n=12

ACF for stationary-resampled comp minus REM timeseries, n=13

ACF for stationary-resampled comp minus REM timeseries, n=14

ACF for stationary-resampled comp minus REM timeseries, n=15

ACF for stationary-resampled comp minus REM timeseries, n=16

ACF for stationary-resampled comp minus REM timeseries, n=17

ACF for stationary-resampled comp minus REM timeseries, n=18

Responses to Simon Chabrillat's Comments

Simon's comments are given in *black italics* and our responses in blue.

General comments

The authors provide a thorough and detailed intercomparison between five modern reanalyses of several diagnostics derived from polar lower stratospheric temperature and vorticity. This should go a long way towards assessing the reanalyses' representation of the potential for polar processing and ozone loss. Following the same approach as a previous paper comparing only ERA-I and MERRA, the focus is on the intercomparison between the reanalyses rather than a comparison with independent observations.

The text is well written with interesting and relevant information provided about the reanalyses and the outcome of the intercomparisons. The figures have been designed with great care to allow overviews of general tendencies as well as comparisons between the reanalyses on specific years. This should allow quick selections of winters with typical (or unusual) polar processing conditions while taking into account the agreement between the reanalyses (or lack thereof), which could be one of the major benefits of the paper. Unfortunately four basic questions are raised here about the methodology used for this comparison (see the major comments below, which are ordered by decreasing importance). It is possible that comments 2.2, 2.3 and 2.4 will be satisfactorily answered with a few sensitivity tests and deeper justifications. Yet I do not see how the choice of MERRA-2 as reference dataset can hold on close examination (for details see comment 2.1 below).

Indeed this manuscript shows that in the polar lower stratosphere MERRA-2 behaves quite differently from all other reanalyses before 1999. This result is important and should be highlighted in the abstract; several previous S-RIP papers reached similar conclusions in the tropical lower stratosphere. But it also invalidates the choice of MERRA-2 as a reference dataset to compute all means and standard deviations of the differences with the four other reanalyses. The choice of the average across four or five reanalyses would provide much additional information on the mean disagreements between the reanalyses and potentially on the variability of these disagreements. This additional information could very well change the conclusions of the study. Hence I believe that it is necessary to re-run at least all the difference diagnostics with another reference dataset and to update the methodology, discussion and conclusions accordingly.

We thank Simon for his thoughtful and detailed comments. Two major changes to the paper motivated by his comments are to use a reanalysis ensemble mean (REM) as a reference for the comparisons rather than using MERRA-2 (see our detailed response below), and to remove MERRA from the reanalyses evaluated in this paper. There are numerous reasons for removing the MERRA comparisons, including the following: The choices that were made by GMAO of which products to archive for MERRA have made "fair" comparisons difficult to impossible for many products, including potential vorticity (PV), which is critical for stratospheric vortex and many other studies. While comparing MERRA with MERRA-2 and other reanalyses was critical to evaluating MERRA-2, numerous such studies have now been done; MERRA-2 was intended to supercede MERRA and sufficient evaluation of it has been done now to warrant this. Finally, especially when using the REM as a reference, it is somewhat problematic to include two reanalyses based on nearly the same model in a comparison of just five reanalyses.

Because these two major changes, especially the switch to using the REM, necessitated a nearly complete rewrite of large portions of the text in the results section (though the final conclusions and relationships among the reanalyses don't change), several of the reviewers' comments now refer to text that has been replaced, and it is not possible to document every change in detail.

Major comments

2.1

Choice of MERRA-2 as a reference dataset

Rather than using a Reanalysis Ensemble Mean (REM), the authors chose MERRA-2 as reference dataset to evaluate the differences between the reanalyses (shown on figs. 5 and 6, 8 and 9, 11 and 12, 14 and 15). In section 2.2.3 the rejection of a REM is first explained by the small size of the sample (5, or 4 if MERRA is excluded due to its expected similarity with MERRA-2) as this "can make the ensemble average sensitive to outliers". Yet this reasoning does not hold if the chosen reanalysis delivers itself many outlying diagnostics. Looking at the mean differences with CFSR, ERA-I and JRA-55 (and even MERRA in several cases; see left columns of figures listed above) the authors correctly note that before 1999 there are very similar band structures located in approximately the same layers and the standard deviations of the differences are remarkably similar in all four reanalyses. This indicates a posteriori that MERRA-2 is not an appropriate choice since it obscures the differences between the four other reanalyses.

In section 2.2.3 it is also explained that "comparing with an average across the reanalyses obscures the actual scale of differences among the individual datasets". Indeed it is quite desirable to show the spread across the reanalyses for each diagnostics, but this consideration should not influence the choice of the reference dataset. The spread can simply be shown by the difference between the maximum and the minimum value reached at each pixel for the considered diagnostic, or alternatively by the standard deviation of the 5 diagnostics (as done by Long et al., 2017, using only three reanalyses).

Since no REM was computed in this study, the authors are left with the difficult choice of one specific reanalysis as reference dataset. The justification for picking specifically MERRA-2 is only given in the conclusions: MERRA-2 is the most recent reanalysis (p.19, l.30). In the absence of independent observations it can be helpful to use such an arbitrary criterion but the selected reanalysis could turn out to be invalid as reference dataset due to special features (or even errors). Hence it is required to check a posteriori that the MERRA-2 diagnostics are inside the range provided by the other reanalyses. As discussed throughout the text this is precisely not the case - especially when one excludes MERRA which is an earlier version of the same reanalysis system and may have pre-processing issues (see next comment). Besides cursory examination of the figures listed above, many sentences in the manuscript highlight this fact (e.g. p.11, lines 32–33; p.12, lines 32–33; p.13, lines 28–29; p.14, line 24; p.15, lines 13–14 *and 26; p.19 line 3) and the summary as well (section 4.1).*

In the tropical regions, two earlier studies have shown that MERRA-2 does not represent correctly the Quasi-Biennal Oscillations before 1995 (Coy et al., 2016, doi:10.1175/JCLI-D-15-0809; Kawatani et al., 2016, doi:10.5194/acp-16-6681-2016). The present manuscript also mentions investigations in progress (Long et al., in preparation) showing that in the SH, the vertical profiles of temperature differences with sondes deliver vertical oscillations in one direction with ERA-I, in the other direction with MERRA-2 and no oscillations with the three other reanalyses (p.11-l.30 to p.12-l.4). These informations cast additional doubts on the pertinence of choosing MERRA-2 as reference dataset.

To summarize, I suggest to use the REM as reference dataset (also in the mean annual cycles in fig. 4, 7, 10, 13) and possibly to replace the standard deviations of the differences by one plot showing the spread of the diagnostics (or their standard deviations). The figures about inter-reanalysis (dis)agreements would show 5+1 plots instead of 4+4. Since the MERRA-2 diagnostics often fall outside the range found with other reanalyses, one could also try to highlight this with some simple line plots.

We have switched to using the REM as the reference dataset, but we kept our “pixel plots” in the same format as before. We feel that our change to using the REM is overall beneficial, because while our results did not change much, this change demonstrated that none of the reanalyses are outliers (to be fair, we were also guilty of using this language when we said that the REM could be sensitive to outliers). In a sample size of 4 to 5, one cannot make judgments about the suitability of a reanalysis as a reference unless there is evidence of it being egregiously different, and using the REM here has shown that to not be the case for any of the reanalyses.

2.2

Derivation of Potential Vorticity in the case of MERRA

Many diagnostics are derived from temperature and Potential Vorticity (PV) on model levels and interpolated afterwards to isentropic levels. In the case of PV this raises special difficulties because only MERRA-2 provides it directly on model levels. Section 2.2.1 explains the preparation of this field in the four other reanalyses:

- *ERA-I : PV is derived from absolute vorticity, T and p on model levels*
- *CFSR : PV is derived from relative vorticity, T and p on model levels*
- *JRA-55 : PV is derived from horizontal wind fields, T and p on model levels*
- *MERRA : PV is read on 42 pressure levels and interpolated back to model levels.*

Hence the MERRA diagnostics on isentropic levels are the result of two successive vertical interpolations, which is not the case for any other reanalysis. These diagnostics are afterwards differentiated numerically w.r.t. equivalent latitude (for MPVG; see p. 9, lines 14–16) or determined from offset criteria (for Sunlit Vortex Averages, see for vortex decay dates, see p. 18, lines 25–28). The final quantities shown (i.e. differences between MERRA and MERRA-2 in figs 11 and 12, 14 and 15, 20 and 21) could turn out to be quite sensitive to this numerical Issue.

Section 4.3 discusses this issue and correctly states that it is important to treat all reanalyses as fairly and equally as possible to reduce the uncertainty in sources of differences. As all five reanalyses, the MERRA dataset includes winds, T and p on model levels. Hence MERRA can easily be pre-processed in exactly the same manner as JRA-55, solving this issue once and for all (this approach could be applied to all five reanalyses to ensure strictly fair comparisons; yet in the experience of this reviewer, such pre-processing details are importantly mainly with respect to the vertical grid hence threaten only the results obtained with MERRA). The fact that this issue is raised with MERRA is especially unfortunate because it is the reanalysis most similar to MERRA-2 which has itself been picked as reference dataset (see previous comment). There is a distinct possibility that a more consistent pre-processing would remove many differences between MERRA and MERRA-2. This would leave us with the expectable (hence unsatisfactory) conclusion that the reanalyses should be grouped between MERRA and MERRA-2 one one side and CFSR, JRA-55 and ERA-I on the other side.

The unavailability of PV or vorticity on the model grid in MERRA has been a hindrance to scientific and comparison studies since its production. In addition to the different calculations Simon mentions here, the PV that is available from MERRA is from the 'ANA' rather than the 'ASM' fields, but the latter are the recommended ones for most purposes (e.g., Fujiwara et al., 2017). The is one of several reasons (see above for others) that we have chosen to remove MERRA from the reanalyses evaluated in this paper.

Even without MERRA, the non-uniformity in PV fields available from different reanalyses is a concern in numerous studies, as we have discussed in the implications and recommendations in our conclusions. A detailed study of differences in PV fields, and the effects of calculating them differently, is obviously beyond the scope of this paper (though it would be valuable and we are initiating such a study). However: (1) the differences in PV fields are most important here for the diagnostics that rely on identifying the vortex edge, and we have made this process more uniform for all the reanalyses as detailed in our response to Simon's major comment 2.4 below; and (2) in response to Simon's last point above, along with other diagnostics, when compared to the REM, it is apparent that MERRA-2 (and MERRA, though no longer shown in the paper) is not a consistent outlier in these diagnostics.

2.3

Analysis based only on daily fields valid at 12-UT

Section 2.1 (p.5, lines 13–14) states that "All analyses are done using daily 12-UT fields from each reanalysis dataset". Are diurnal cycles completely negligible, even for diagnostics like Tmin when sunlight comes back? This seems like a serious assumption to me considering longitudinal asymmetries (look e.g. at Fig.5 in Lawrence et al., 2015) and also the real possibility that such diurnal cycles could be larger in some reanalyses than in other ones.
A simple way to check this assumption would be to re-run the diagnostics using e.g. only 0-UT fields. If the results turn out very similar, this sensitivity test would still be worth mentioning in the text. But any diagnostic showing non-negligible dependence on time of day (i.e. not the same results using 0-UT fields than 12-UT fields) should be run using 6-hourly

fields as input.

We re-ran all of our diagnostics using 00UT, and repeated our analyses -- everything came out virtually identical. We now mention in the paper that we have tested using 00UT data and that it does not affect our results.

2.4

Definition of the vortex edge

The discussion on Sunlit Vortex Area (p.16 lines 1–2) is not quite clear. I understand it as

follows: “Investigation of the reanalyses’ differences in total vortex area from MERRA-2 reveals that they are nearly identical to the ones for SVA. This indicates that the SVA differences from MERRA-2 are largely dominated by differences in the size of the vortex edge contours area rather than vortex shape.”. If this is correct, one wonders why the manuscript does not simply show, compare and discuss the vortex areas themselves.

Of course the determination of the vortex area closely depends on the method chosen to define the vortex edge. Here it is determined directly from the sPV values, using a constant vertical profile of sPV limits as a function of potential temperature (p. 9, lines 21–24 and Lawrence and Manney, 2017). This approach makes sense when using a single reanalysis, e.g. to interpret observations and their variations in time. Yet I wonder if this is valid when comparing several reanalyses because (contrarily to equivalent latitudes) they may have different ranges of sPV on any given day and isentropic level. It seems to me that the the classical vortex edge definition (equivalent latitude of the maximum of the wind speed times the PV gradient: Nash et al., 1996; Manney et al., JGR, 2007) could be more appropriate to the problem at hand because it would adapt itself to the different ranges of PV potentially delivered by each reanalysis. This could also deliver more robust evaluations of the vortex decay dates (p.19, lines 22–27).

The Nash method and the windspeed*PV gradient methods are slightly different (see, e.g., Manney et al., 2007, JGR for a discussion of this), but the main issue is that using any method to get daily varying vortex edge values would complicate and possibly contaminate the intercomparisons. Daily methods are prone to giving spurious jumps and oscillations that can dramatically change the vortex edge value (and thus quantities such as vortex area) from one day to the next. If this happened at different times in different reanalyses, the results could be unnecessarily skewed. As examples of this behavior, consider the vortex area and vortex edge quantities provided on the ozonewatch.gsfc.nasa.gov website, which catalogs these quantities for MERRA-2 data (which we have confirmed are based on the Nash method).

The following is vortex area (in million km^2) and vortex edge PV (as “modified” PV) for the currently ongoing SH winter at 460K:

Date	Area (km^2)	VortEdge MPV
2018-07-09	18.90	-27.15
2018-07-10	23.76	-25.00
2018-07-11	29.05	-22.80
2018-07-12	35.04	-20.26
2018-07-13	39.77	-18.30
2018-07-14	45.01	-16.17

In this case, the vortex apparently grows nearly 5 million square kilometers per day (roughly 2% of a hemisphere per day), from an all time climatological minimum vortex area to above 90th percentile (for this level), all in the span of 5 days. Examination of PV maps does not support there being a significant change in the vortex area at this time
[see the plot at
https://ozonewatch.gsfc.nasa.gov/meteorology/figures/merra2/pv/mpvweas_460_2018_merra2.pdf and the data from
https://ozonewatch.gsfc.nasa.gov/meteorology/figures/merra2/pv/mpvweas_460_2018_merra2.pdf and
https://ozonewatch.gsfc.nasa.gov/meteorology/figures/merra2/pv/mpvwes_460_2018_merra2.txt;
PV maps can be viewed at https://acd-ext.gsfc.nasa.gov/Data_services/antarctic/history.html].

Note that cases such as this one are not unique; other examples in other years, at other levels, and in both hemispheres are relatively common. For these reasons (also see the discussion in Manney et al., 2007, JGR, regarding the robustness of daily varying versus constant vortex edge values), we do not like to use daily varying vortex edges and prefer instead to use constant vortex edges. Although using constant vortex edges is also unrealistic in the sense that a single PV value on a single isentropic level cannot be the vortex edge forever, as long as an appropriate value is chosen within the strong PV gradient region near the vortex edge, the chosen contour will grow and decay in a realistic and more gradual manner that is still representative (at some times more so than a time-varying value) of the vortex.

We have switched to using constant vortex edge values that were calculated for each reanalysis individually, using the climatological average PV values at the maximum PV gradients. We chose to strictly use PV to determine the edges in this case so that we would not introduce another field/quantity (winds or otherwise), with its own uncertainties, into the mix.

2.5
Need to add some auxiliary information
Repeatability of these results requires several pieces of methodological information to be provided explicitly, e.g. in an annex or in a supplement:

- *Last paragraph in section 2.2.1: please list of the 13 pressure levels used in the paper and their “climatologically corresponding isentropic surfaces”.*

We now list the pressures and potential temperatures of the levels (there are actually 14 including both “boundary” ones) that are used in the appendix.

• P.9, line 2: please provide the climatological profiles of H2 O and HNO3 used to determine TICE and TN AT (with original reference if available).

A table of the values and paragraph on how they are calculated have been added as an appendix.

• P.9, line 21–24: please list the sPV values defining the vortex edges at each isentropic level (unless of course the vortex edge definitions is changed - see previous comment).

Because we now calculate the vortex edge sPV values used as a function of altitude for each reanalysis separately (as described in response to 2.4), a new figure has been added that shows the profile of vortex edge sPV values used for each reanalysis. As noted in the “data availability” text, our diagnostics (including the vortex edge profiles used) are available by contacting the authors.

2.6

Opportunity to illustrate disagreements on a specific winter season

One of the main goals of S-RIP is to increase in the community the awareness about the uncertainties hidden in the reanalyses, not only w.r.t. inter-annual variations but also for case studies. All diagnostics are shown as yearly time series which is useful to quickly evaluate the level of agreement between the reanalyses on any given winter. This provides an opportunity to highlight the potential disagreements between these diagnostics on a specific winter (and hemisphere) chosen for that purpose. So I suggest to select a diagnostic, year and level which highlight such disagreements between the reanalyses, and to plot this diagnostic either with line plots (e.g. time variations during that winter) or with maps (e.g. five maps for one specific date). Such an illustration would be easy to realize and could improve the impact of the paper. For example, Figure 15 indicates that even on some recent years the difference of SVA between MERRA (or CFSR) and MERRA-2 can be as large as 2% of the NH, over vortex areas which have seasonal averages of around 8%. That seems quite large and may warrant a few detailed maps (unless of course this outcome does not hold after examination of the previous comments).

While an examination of differences in case studies among the reanalyses would be an interesting and potentially valuable study, examining just one case would provide an incomplete and possibly misleading impression of the detailed reasons for and morphology underlying the differences we show. The differences and reasons for them will vary by hemisphere, by the time period chosen because of different data inputs into the reanalyses, and, especially in the NH, because of interannual and intraseasonal variability in meteorological conditions (e.g., for disturbed versus quiescent meteorological conditions). Because this type of analysis would be worth a more detailed study on its own where sensitivity to such conditions could be explored, and because our paper was already too long, we consider this beyond the scope of this paper.

3

Minor comments and corrections

- *Abstract: long enough that it would be useful to split it into two or three paragraphs.*

We have divided the abstract into several paragraphs (after modifying it to reflect the changes in the paper) to more clearly separate different results highlighted.

- *P.1, lines 16–17: "Some reanalyses show convergence toward better agreement in vortex diagnostics after 1999, while others show some persistent differences across all years" - please be specific, i.e. identify which reanalyses agree better and which ones have persistent differences.*

This text has been modified to reflect the patterns seen in differences from the REM, so the text in question has been removed.

- *P.1, lines 24–25: "the large interannual variability of NH winters has given rise to many seasons with marginal conditions and high sensitivity to reanalysis differences". Please re-phrase to be more precise: this sensitivity is clearly seen for the vortex decay dates (fig. 21) but not for the number of days (fig. 17) and fraction of vortex volumes (fig. 19) with $T < TN AT$.*

We have reworded this sentence to indicate that the results are more sensitive in the NH and that this is particularly apparent in the vortex decay dates.

- *P.2, lines 21–22: or more correctly, "detection of recovery from chemical ozone depletion also requires accurate knowledge of variability and long-term changes in polar vortex dynamics and temperatures."*

We have changed this as suggested.

- *P.2, line 23: "the conversion of chlorinated species into forms...". Please also mention brominated species.*

We have changed the wording as suggested and added "and brominated".

- *P.2, line 29: Please define "active chlorine" (i.e. Cl and ClO)*

We now spell it out: Cl+ClO+2ClOOCl

- *P.4, lines 5–6: "...much larger temperature biases for NCEP/NCAR than in the other* *reanalyses, not only because that reanalysis is unsuitable for stratospheric studies..." . This looks to me like a confusion between cause and effect. Consider writing instead something like "...not only due to shortcomings of the former, but..."*

We have reworded this along the lines suggested.

- *P.4, line9: CFSR is not considered as a "full-input reanalysis"? Why?*

This mention of "full-input reanalysis" was within a parenthetical where we were explaining JRA-55 as "the Japan Meteorological Agency's latest reanalysis assimilating both surface and upper air observations"; we used this same parenthetical to explain that a reanalysis that assimilates both surface and upper air observations is considered full-input. We did not intend for it to sound as if CFSR was not a full-input reanalysis. While we did/do refer to CFSR/CFSv2 as a full-input reanalysis elsewhere, we understand that this text is confusing, so we have moved the full-input description elsewhere to be more clear.

- *P.4, line 26: "... during much of which ..." - please re-phrase*

We rephrased this sentence.

- *P.4, line 16–28: It seems to me that temperatures profiles retrieved from limb-scanning instruments (UARS-MLS, HALOE, SABER, MIPAS, Aura-MLS, ACE-FTS) could provide a valuable source of independent data to evaluate the reanalyses. Have such comparisons been done already for some reanalyses (or NWP analyses) with a focus on the polar lower stratosphere? If yes, the corresponding papers should be cited in the introduction. If no, this could mean that those instruments are not fit for this purpose and this warrants a short explanation in the introduction (e.g. due to lack of precision? lack of accuracy? lack of horizontal resolution?).*

The introduction already had two paragraphs describing previous studies comparing polar processing diagnostics with observational data (in the discussion paper, page 3 line 32 through page 4 line 28). In addition, part of one of those paragraphs discusses why the diagnostics we compare here cannot be easily compared with observations (in the discussion paper, page 4, lines 21 through 26). We have, however, expanded a bit in the first of those paragraphs on why it is difficult (sometimes impossible) to compare polar processing diagnostics with limb-sounding data.

- *P.4, lines 31–32: the history of the availability of specific reanalyses on native model levels is a quite technical matter for the introduction. I think that this sentence can be removed (such considerations are well explained in the next section).*

We have deleted this sentence as suggested.

• P.5, lines 14–17: Sentence is too long cumbersome (consider splitting). Replace "...importance of resolution, especially the vertical grid,..." with "...importance of resolution, especially in the vertical dimension,...".

We replaced "the vertical grid" with "in the vertical dimension" as you suggest. As we no longer include MERRA in the mix the "except where unavailable (e.g., PV for MERRA)" bit got deleted. We think and hope the sentence reads fine now.

• P.5, line 19: define acronym "GMAO" or maybe drop it (anyway GMAO is the only division delivering atmospheric reanalyses at NASA).

GMAO is now defined.

• P.5, lines 23–24: I think that this approach is not valid, and that PV should be recomputed from u, v, T, p on model levels as you did for JRA-55 (see major comment above).

As noted in the major changes and the response to major comment 2.2, we have removed MERRA from the intercomparison, so this text has been removed.

• P.6, line 8: "...but a much older earlier version than that used in MERRA-2 ".

The text read: "a much older version". This sentence was deleted because we no longer analyze MERRA.

• P.6, line 15: Please replace this URL by a proper bibliographic reference

We have reformatted the citation to give the URL in a proper bibliographic reference.

• Section 2.1.2: the distinction between CFSR and CFSv2 is difficult to follow. The title of the section should be changed to the full name of the dataset, i.e. "CFSR/CFSv2". I advise to explain the distinction upfront: "NCEP-CFSR/CFSv2 (hereinafter CFSR) (Saha et al., 2010) is a global reanalysis covering the period from 1979 to the present 2010. From 2011 onwards it is superseded by CFSv2 (Saha et al., 2014)." and to finish the subsection with a simple sentence about the naming convention, e.g. "Hereafter CFSR/CFSv2 is designated simply by CFSR".

Consistent with recently clarified recommendations from the S-RIP project, we now use "CFSR/CFSv2" throughout the paper. We have clarified the text regarding the transition between the two.

• Figure 1: Consider ending the caption with "See Fujiwara (2017, Fig.8) for a similar time line but organized per instrument." (this is only a suggestion).

Done.

• Section 2.2.1: Consider citing also Manney et al. (JGR, 2007) which provided an excellent overview about the Derived Meteorological Products used here.

The Derived Meteorological Products described by Manney et al (2007) are not used in this paper. The PV scaling used here was discussed in that paper, but more thoroughly in the earlier papers that we already cite.

• Figure 2 and first paragraph of section 3: I do not think that these are really useful (they explain obvious concepts). Consider removing. If you decide to keep, line 23: replace "To demonstrate..." by "To illustrate...".

The original Figure 2 and corresponding text have been removed considering both your suggestion and that of the other reviewer to do so.

• P.11, lines 19–21: AIRS was introduced on September 2002 in MERRA and MERRA-2 and on February 2003 in ERA-I and CFSR. The potential importance of this instrument is an interesting point which deserves a few more details. For example, add a reference about its sensitivity to stratospheric temperatures. Similarly interesting comments could be added about GPS-RO which saw assimilation into ERA-I and CFSR from 2001, into MERRA-2 from 2004, into JRA-55 from 2006 and never into MERRA.

We are now more specific about the introduction of AIRS data in different reanalyses and added a reference (Hoffman and Alexander 2009) that demonstrates AIRS' sensitivity to stratospheric temperature. We believe that ERA-I and CFSR/CFSv2 started assimilating AIRS in 2004, not 2003. MERRA-2 actually began assimilating AIRS in September 2002, not in 2003 as we mistakenly stated. We added two sentences about GNSS-RO data, including the years when they were assimilated and impact. We cite Fujiwara et al. 2017 for details.

• P.12, line 30: "...(as was the case in the SH) they remain larger in CFSR..."

Done

• P.12, line 35: "...a clear decrease in them is not evident". Please re-phrase.

This sentence was eliminated in the revised manuscript.

• Discussion of figure 8 and 9 (p.13 line 26 to p.14 line 13): the standard deviations of the differences are not discussed at all, even though striking differences can be sen with the corresponding Tmin diagnostics in both hemispheres (i.e. compare right columns of fig.5 with fig.8 and fig.6 with fig.9). If you decide to keep showing standard deviations of

differences (see first major comment) it would make sense to discuss this.

We agree. A discussion of standard deviations in A_NAT is included in both the SH and NH paragraphs. The final paragraph of this subsection notes the overall agreement and discrepancies between the patterns of statistical significance and standard deviations between the two diagnostics and ascribes them to minor differences between reanalyses in the morphology of the fields. The latter point was already made in the original manuscript but only for the SH.

- *P.14 line 29: remove words "... show that the variances of the differences..."*

Done. The sentence now reads: ""The standard deviations tend to increase with height (...)

- *P.15 line 6: see corresponding major comment (2.3)*

As per our discussion of the major comment, the MERRA comparisons have been removed from the paper, so we deleted this text.

- *P.15 line 37: "... (indicating that these reanalyses haveing higher larger SVA than MERRA-2)..."*

This text has been modified due to the switch to using the REM rather than MERRA-2 as a reference.

- *Section 3.3 is very long and tedious to read, probably because it includes the methodology about the diagnostics shown here. Consider moving this methodology to a new subsection in section 2.*

We have moved the paragraphs describing the methodology for these diagnostics to a new subsubsection within section 2.

- *Figures 16–17 and P.19 line 19: It looks like summing the number of days over lower stratospheric levels implies a close dependence on the vertical pressure grid used for this diagnostic. Is there a way to avoid this? In any case the explicit listing of these pressure levels is even more necessary (see major comment 2.4).*

This diagnostic does inherently depend on the resolution and number of pressure levels. There are probably good ways to remove this dependence (e.g., by taking a column average number of days, or selecting the column maximum), but we have removed (what were formerly) Figures 16 and 17 and their discussion to help shorten the paper as requested by reviewer 1, and because the original Figures 18 and 19 provided much the same information in the context of the intercomparisons.

- *Figures 16–19: I understand that plots (b) and (d) simply show the same sensitivity range as already shown by the bars in plots (a) and (d) but zoomed and centered on zero? If this is wrong, the captions and text require clarification. If this is right, the usefulness of plots (b) and (d) is not clear since they could be removed while not changing the discussion of the figures (also because the figures show sensitivities which do not depend much on the reanalyses nor on the year).*

Your first interpretation is correct; that is, panels b and d show the same sensitivity ranges, but centered on zero. We include them because the human eye has trouble properly assessing the lengths of bars/lines when they are in different contexts, such as being centered at different heights as in panels a and d (see also, e.g., the Ponzo illusion and/or the Müller-Lyer illusion). We have kept these panels because we do discuss them specifically (however, we have made these references more apparent), and we think it is important to show how the sensitivities compare among the reanalyses.

- *P. 17 line 19: After 1999, it looks like fig. 17c has quasi-biennal periodicity. Could there be a link between this diagnostic and the QBO?*

We do plan to look more in depth at potential links between the QBO and polar processing diagnostics with some other S-RIP collaborators, but for now, this is beyond the scope of the paper.

- *P.17 line 21: Why is the important diagnostic VPSC/Vvort and not VPSC itself? If possible, explain this in one additional sentence.*

We have reworded this and added a phrase to note that this scaling is used to get a diagnostic that is independent of the (considerable) interannual and interhemispheric variations in vortex size.

- *P.17 lines 27–28: "...with a range among the reanalyses of 0.98 – 1.30 km..."*

This is not accurate because the altitude differentials are independent of the reanalyses. The Knox approximation is a simple approximation to go from potential temperature to altitude. Since we are vertically integrating areas to get volumes, we need the altitude widths of the potential temperature levels (i.e., the altitude differentials, or thinking in terms of integrals, the "dz" values). We have clarified the text to make this more explicit that we are referring to the minimum and maximum altitude differentials.

- *P.17 lines 28–29: I expect that the winter mean is applied at the end, i.e. you discuss winter means of daily fractions rather than the fractions of winter mean volumes? Please clarify, if possible in the caption of fig. 18 as well.*

Yes, this is correct. We have clarified the text to make this explicit.

• *P.18 lines 3–4: this last sentence is not useful (see also comment above about plots (b) and (d) not useful in these figures).*

We have reworded this sentence to make it clearer that we consider this worth saying as it is evidence against persistent biases in horizontal temperature gradients. Also see our response to your comment on Figure 16-19 above.

• *P.18 lines 11–12: "...with differences betweeen reanalyses indicating some differences in horizontal temperature gradients (especially in, e.g., 2011 and 2014).". Can this be seen directly on fig. 19 or are you commenting figures which are not shown in the manuscript? Please clarify.*

Persistent biases in the temperature gradients would result in different sensitivity to the changes in PSC threshold values used. In addressing the previous comment, we have revised the text to make this more explicit.

• *Figures 20–21: please align the titles of the figures with the vocabulary used in the text (i.e. vortex decay dates - not vortex breakup dates). Since the ranges are the same for both figures, it is possible to clarify the caption: "...differences that greatly exceed the range by more than 7 days larger than 21 days are marked with a white X."*

The figure titles and captions have been modified along the lines requested.

• P.18 line 35: "For the SH, Figure 20a shows that..."

Done.

• *P.19 line 8: It is possible to get away from the figure and closer to its meaning. For example: "Fig.20 also shows that on a few years (such as 2002 and 2009) the vortex decayed at a much later date in MERRA, JRA-55 and CFSR than in MERRA-2.*

This text was changed in switching from looking at differences from MERRA-2 to differences from the REM. We have tried to word all this discussion in terms of earlier/later vortex decay dates rather than positive/negative differences.

• *P.20 line 7: It is not fair to compare agreement generally within about 1K after 1998 with disagreements up to about 6K before.*

The numbers have been altered by using the REM, but we have reworded the corresponding sentence to specify the range of differences in a consistent way in each case.

• *P.22 line 34: "...and after rigorous assessment of the relationships of temperature changes* *to observations assimilated (which, to our knowledge has not been done)." It could be that this has not yet been done systematically due to the huge diversity of assimilated observations. Yet Simmons et al. (QJRMS, 2014, doi:10.1002/qj.2317) already provided a remarkable first step in this direction.*

Simmons et al. did, indeed, provide a fairly detailed analysis of responses to inputs for ERA-Interim, and compared the results for long-term variations with that in ERA-40 (now obsolete), MERRA (now becoming obsolete), and JRA-55. They did not do such a detailed analysis of the other reanalyses, nor include MERRA-2 (not yet available at that time) or CFSR/CFSv2. We had overlooked mentioning this study in the introduction when we discussed previous intercomprisons, an omission that we have remedied. We have qualified the statement here to read "..has not been done for most of the reanalyses considered here."

- *P.23, last sentence: "...the comparison of reanalyses is a powerful tool for assessing robustness and uncertainty in these diagnostics." Yes but this is still an incomplete tool because the reanalyses often use similar parametrizations and assimilate very similar observational datasets.*

This point is, indeed, made in several places in the text of the paper. We feel this statement (which does not imply that it is the only powerful tool or that it is a perfect tool) is appropriate for closing the text of the paper.

Reanalysis intercomparisons of stratospheric polar processing diagnostics

Zachary D Lawrence[1,2], Gloria L Manney[2,1], and Krzysztof Wargan[3,4]

[1]New Mexico Institute of Mining and Technology, Socorro, NM USA
[2]NorthWest Research Associates, Socorro, NM USA
[3]NASA/Goddard Space Flight Center, Greenbelt, MD USA
[4]Science Systems and Applications Inc., Lanham, MD, USA

Correspondence: zachary.lawrence@student.nmt.edu

Abstract.

We compare herein polar processing diagnostics derived from the ~~five~~ four most recent full-input reanalysis datasets: the National Centers for Environmental Prediction Climate Forecast System Reanalysis ~~(CFSR~~/ Climate Forecast System, version 2 (CFSR/CFSv2), the European Centre for Medium-Range Weather Forecasts Interim Reanalysis (ERA-Interim), the Japanese Meteorological Agency's Japanese 55-year Reanalysis (JRA-55), and the National Aeronautics and Space Administration's Modern Era Retrospective-analysis for Research and Applications version ~~1 (MERRA) and version~~ 2 (MERRA-2). We focus on diagnostics based on temperatures and potential vorticity (PV) in the lower to middle stratosphere that are related to formation of polar stratospheric clouds (PSCs), chlorine activation, and the strength, size, and longevity of the stratospheric polar vortex.

Polar minimum temperatures (T_{min}) and the area of regions having temperatures below PSC formation thresholds (A_{PSC}) show large persistent differences between the reanalyses, especially in the southern hemisphere (SH), for years prior to 1999. Average absolute differences ~~between the reanalyses~~ of the reanalyses from the reanalysis ensemble mean (REM) in T_{min} are as large as ~~6~~ 3 K at some levels in the SH (~~2~~ 1.5 K in the NH), and absolute differences ~~in~~ of reanalysis A_{PSC} ~~larger than 2~~from the REM up to 1.5% of a hemisphere (~~1~~0.75% of a hemisphere in the NH). After 1999, ~~there is a dramatic convergence~~ the reanalyses converge toward better agreement ~~between the reanalyses~~ in both hemispheres~~throughout the lower stratosphere, with average~~ , dramatically so in the SH: Average T_{min} differences from the REM are generally less than 1 K in both hemispheres, and average A_{PSC} differences less than ~~0.5~~0.3% of a hemisphere.

The comparisons of diagnostics based on isentropic PV for assessing polar vortex characteristics, including maximum PV gradients (MPVG) and the area of the vortex in sunlight (or sunlit vortex area, SVA), show more complex behavior: ~~Some reanalyses show~~ SH MPVG showed convergence toward better agreement ~~in vortex diagnostics~~ with the REM after 1999, while ~~others show some persistent differences across all years~~NH MPVG differences remained largely constant over time; differences in SVA remained relatively constant in both hemispheres. While the average differences from the REM are generally small for these vortex diagnostics, understanding such differences among the reanalyses is complicated by the need to use different methods to obtain vertically-resolved PV for the different reanalyses.

We also evaluated other winter season summary diagnostics, including the ~~number of days below PSC thresholds integrated over vertical levels, the~~ winter mean volume of air below PSC thresholds, and vortex decay dates. For ~~these summary diagnostics~~the volume of air below PSC thresholds, the reanalyses generally agree best in the SH, where relatively small interannual variability has led to many winter seasons with similar polar processing potential and duration, and thus low sensitivity to differences in meteorological conditions among the reanalyses. In contrast, the large interannual variability of NH winters has given rise to many seasons with marginal conditions ~~and high sensitivity~~ that are more sensitive to reanalysis differences. For vortex decay dates, larger differences are seen in the SH than in the NH; in general the differences in decay dates among the reanalyses follow from persistent differences in their vortex areas.

Our results indicate that the transition from the reanalyses assimilating Tiros Operational Vertical Sounder (TOVS) data to Advanced TOVS and other data around 1998 – 2000 ~~data~~ had a profound effect on the agreement of the temperature diagnostics presented ~~,~~ (especially in the SH) and to a lesser extent the agreement of the vortex diagnostics. ~~Our results lead to~~ We present several recommendations for ~~usage of~~ using reanalyses in polar processing studies, particularly related to the sensitivity to changes in data inputs and assimilation. Because of these sensitivities, we urge great caution for studies aiming to assess trends derived from reanalysis temperatures. We also argue that one of the best ways to ~~present~~ assess the sensitivity of scientific results on polar processing is to use multiple reanalysis datasets.

1 Introduction

Past, present, and future polar lower-stratospheric ozone depletion is a subject of critical scientific and human interest. Not only does chemical ozone depletion depend critically on temperatures and polar vortex dynamics in the lower stratosphere, but changes in lower stratospheric ozone also feed back and alter dynamical conditions in both the stratosphere and troposphere, which can significantly affect surface climate (Polvani et al., 2011; Albers and Nathan, 2013; WMO, 2014; Waugh et al., 2015, and references therein). Moreover, ozone depletion is affected by changing tropospheric and stratospheric temperatures, and in turn alters those temperatures via radiative forcing (e.g., Lacis et al., 1990; Forster and Shine, 1997; Levine et al., 2007; Hegglin et al., 2009; Telford et al., 2009; Riese et al., 2012; WMO, 2014). The southern hemisphere (SH) springtime polar vortex breakup disperses ozone-depleted air over populated regions, increasing surface UV exposure (e.g., Ajtić et al., 2003, 2004; Pazmino et al., 2005; WMO, 2007). Our ability to quantify chemical ozone loss in observations and to fully understand the mechanisms resulting in that destruction is a key to improving our modeling capability, which in turn will allow accurate forecasting of future ozone changes and their feedbacks on weather and climate. That ability depends critically on accurate knowledge of temperatures in and the dynamics of the lower stratospheric winter and springtime polar vortices. Even in the Antarctic, interannual variations in chemical ozone loss are controlled largely by variations in polar vortex dynamical conditions; thus detection of ~~ozone recovery~~ recovery from chemical ozone depletion also requires accurate knowledge of variability and long-term changes in polar vortex dynamics and temperatures (e.g., Newman et al., 2004; Huck et al., 2005; WMO, 2014).

The chemistry leading to ozone loss involves the conversion of ~~chlorine~~ chlorinated and brominated species into forms that destroy ozone on the surfaces of cold aerosol particles and/or polar stratospheric clouds (PSCs) (see, e.g., Solomon, 1999; WMO, 2014, for reviews). These processes only occur at temperatures below a threshold that is dependent on pressure and on water vapor (H_2O) and nitric acid (HNO_3) concentrations (e.g., Hanson and Mauersberger, 1988; Solomon, 1999). Furthermore, these processes can only result in widespread and persistent chlorine activation when/where the cold air is confined so that mixing with outside air cannot dilute the activated fields – that is, inside the "containment vessel" of the winter and springtime stratospheric polar vortex (e.g., Schoeberl and Hartmann, 1991; Schoeberl et al., 1992). Finally, the reactions by which active chlorine (Cl+ClO+2ClOOCl) destroys ozone require sunlight. The formation and maintenance of the dynamical and chemical environment described above is referred to as "polar processing" since all of these conditions are required for chemical ozone depletion to take place in the lower stratosphere. Since reanalyses from data assimilation systems (DAS) are among the best available tools for modeling and understanding stratospheric dynamics, as well as for driving models of past and present ozone loss, the representation of lower stratospheric temperatures and vortex dynamics in these reanalyses is critical to furthering our understanding of and ability to predict ozone depletion and eventual ozone recovery.

Over approximately the past two decades, numerous studies have compared meteorological products from DAS in the polar stratosphere (e.g., Manney et al., 1996; Pawson et al., 1999; Davies et al., 2003; Feng et al., 2005), and/or compared such products with observations (e.g., Knudsen et al., 1996, 2001, 2002; Gobiet et al., 2007; Boccara et al., 2008; Tomikawa et al., 2015); see, e.g., Lawrence et al. (2015) and Lambert and Santee (2018) for further review.

One important finding of earlier studies was that the NCEP/NCAR (National Centers for Environmental Prediction / National Center for Atmospheric Research) and NCEP/Department of Energy reanalyses are unsuitable for polar processing studies because of their poor representation of the stratosphere (very low model top and few model levels) and outdated assimilation approaches (e.g., assimilation of retrieved temperature from operational sounders) (e.g., Manney et al., 2003, 2005a, b). The European Centre for Medium-range Weather Forecast's (ECMWF's) 40-year Reanalysis (ERA-40) reanalysis was also shown to be unsuitable for such studies, partly because of unrealistic oscillations in the temperature profiles (e.g., Manney et al., 2005a, b; Feng et al., 2005; Simmons et al., 2005). In the past few years, some studies have begun focusing on the latest generation of reanalyses, which have vast improvements in models and assimilation methods and more comprehensive data inputs (for a review of reanalysis characteristics, see Fujiwara et al., 2017). WMO (2014) showed comparisons of potential PSC volume and of springtime vortex breakup dates (calculated as in Nash et al., 1996) between NCEP/NCAR and two modern reanalyses, ECMWF's Interim reanalysis (ERA-Interim) and the NASA Global Modeling and Assimilation Office's (GMAO's) Modern Era Retrospective analysis for Research and Applications (MERRA); NCEP/NCAR was shown to give much lower PSC volumes than in the more modern reanalyses, and the vortex breakup dates differed substantially among each of the reanalyses. Simmons et al. (2014) provided a detailed analysis of the effects of DAS inputs on long-term variability and trends in the ERA-Interim reanalysis temperatures and made comparisons with MERRA, the Japanese 55-year reanalysis (JRA-55), and the older ERA-40 reanalysis. Lawrence et al. (2015) compared a large suite of diagnostics, based on polar vortex characteristics and temperatures, that are important for polar processing between MERRA and ERA-Interim (hereinafter referred to as ERA-I) for the then-available 34 years of those reanalyses. These comparisons showed significant changes in agreement between the reanalyses over that period, with overall good agreement in the period since 2002, when the amount of data ingested into the two reanalyses' DAS was much greater; the largest improvements in agreement were particularly seen in Antarctic temperature diagnostics. In a paper describing global temperature and wind comparisons as part of the Stratosphere-troposphere Processes

And their Role in Climate (SPARC) Reanalysis Intercomparison Project (S-RIP) (Fujiwara et al., 2017), Long et al. (2017) also emphasized changes in agreement between reanalyses related to data input changes, especially improvements in temperature agreement after the transition from TOVS to ATOVS around 1998 to 2000; they also pointed out issues with discontinuities in some reanalyses that were run in multiple streams. Changes such as these noted by Lawrence et al. (2015) and Long et al. (2017) argue for great hesitancy in using temperatures and other fields from individual reanalyses for diagnosing long-term changes and trends.

The ability to compare polar processing diagnostics with observations is very limited for several reasons. Somewhat paradoxically, the vast improvements in DAS usage of available observations have resulted in there being very few truly independent temperature datasets. Furthermore, many of the datasets that are available, even those ingested into the DAS, generally suffer from very limited spatial and/or temporal coverage (e.g., balloon-borne and lidar measurements) and/or issues with resolution, precision, and length of data records (e.g., limb-sounding research satellites). For example, many of the limb-sounding satellites do/did not retrieve temperatures down far enough to fully cover the lower stratosphere and have very coarse vertical resolution, and those that do typically have incomplete coverage of the polar regions (e.g., Aura MLS does not observe poleward of 82°latitude); further, validation studies generally do not indicate better quality in lower stratospheric temperatures from limb sounders than that of the reanalyses (e.g., Schwartz et al., 2008, and references therein) . Nevertheless, several recent studies have compared some of the latest generation reanalyses with observations: For example, Hoffmann et al. (2017) compared MERRA, MERRA-2 (the recent successor to MERRA), ERA-Interim, and NCEP/NCAR reanalyses with temperatures and winds from long-duration Concordiasi ballon flights in the Antarctic lower stratopshere from September 2010 through January 2011; unsurprisingly, they found much larger temperature biases for NCEP/NCAR than in the other reanalyses, not only because ~~that reanalysis is unsuitable for stratospheric studies~~of the shortcomings in that reanalysis, but also because the other reanalyses they considered assimilated Concordiasi measurements. Lambert and Santee (2018) compared MERRA, ~~MERRA-2, ERA-Interim, JRA-55~~ MERRA-2, ERA-Interim, JRA-55 (the Japan Meteorological Agency's latest reanalysis assimilating both surface and upper air observations~~, hereinafter referred to as a "full-input" reanalysis~~), and ~~CFSR~~ CFSR/CFSv2 (referring collectively to NCEP's Climate Forecast System Reanalysis and Climate Forecast System Version 2) with COSMIC (Constellation Observing System for Meteorology, Ionosphere and Climate) GPS-RO (Global Positioning System–Radio Occultation) temperatures, and presented an innovative analysis using thermodynamic calculations to derive an independent temperature reference from satellite observations of HNO_3, H_2O (from the Aura Microwave Limb Sounder, MLS), and PSC aerosols (from Cloud-Aerosol Lidar with Orthogonal Polarization). They found temperature biases in the reanalyses with respect to COSMIC of -0.6 to +0.5 K, and biases ranging from -1.6 to +0.1 K with respect to the derived temperature references for two PSC types.

The use of multiple data sources and novel methods allowed Lambert and Santee (2018) to compare temperatures over a wide range of winter polar vortex conditions in both hemispheres for 2008 through 2013. Studies comparing with other data sources, such as long-duration balloon flights (Hoffmann et al., 2017, and references therein), are generally restricted to more limited spatial and temporal regimes. In addition, many of the latest generation reanalyses assimilate data sources such as COSMIC GPS-RO and long-duration balloon flights (Fujiwara et al., 2017; Hoffmann et al., 2017; Lambert and Santee, 2018), thus complicating interpretation of differences from those data sources. Also, as noted by Lawrence et al. (2015), some of the most useful diagnostics of polar processing, while conceptually simple, depend on having full and dense coverage of the polar regions (e.g., minimum high-latitude temperatures or area of temperatures below a PSC threshold), and/or are based on vortex diagnostics that are defined by potential vorticity (PV) (e.g., vortex area or vortex-edge PV gradients) that do not have corresponding observations. Furthermore, reanalyses are used in polar processing studies that span the 35 (or more) years of their duration, ~~during much of which there are no data available~~ but much of this period lacks data with widespread coverage for comparison. Because of these limitations, comparisons of reanalyses remain one of our most valuable tools for assessing their representation of the dynamical conditions that control polar chemical processing and ozone loss.

Since the work described in Lawrence et al. (2015), the ~~MERRA-2 reanalysis~~ MERRA-2 reanalysis (intended as a replacement for MERRA) has become available and widely used, including for polar processing and polar vortex studies ~~(e.g., Manney and Lawrence, 20~~ ~~Since that work was completed, the CFSR reanalysis was made available on the native model levels, and we have also obtained and processed the complete JRA-55 record.~~ (e.g., Manney and Lawrence, 2016; Lambert and Santee, 2018; Lawrence and Manney, 2018). While Long et al. (2017) compared temperatures in all of the latest generation reanalyses, they focused on zonal means and the whole stratosphere rather than on the polar lower stratosphere and diagnostics specifically relevant to polar processing. To our knowledge, no studies have been done that compared lower stratospheric polar processing diagnostics in ~~all five recent full input reanalyses (MERRA, MERRA-2, ERA-I, JRA-55, and CFSR~~the four most recent “full input” (a dataset that ingests both surface and upper air observations) reanalyses (MERRA-2, ERA-I, JRA-55, and CFSR/CFSv2).

In this paper, we compute and analyze the diagnostics used by Lawrence et al. (2015) to provide a more complete and quantitative characterization of reanalysis differences during the satellite era. In addition to including the ~~MERRA-2, JRA-55, and CFSR~~ MERRA-2, JRA-55, and CFSR/CFSv2 reanalyses, the calculation and analysis of diagnostics has been updated to include sensitivity tests (e.g., to temperature ~~and vortex edge~~ thresholds) and to include assessment of the variability in reanalysis differences and the statistical significance of those differences. Section 2.1 briefly describes the reanalysis datasets and the assimilation system inputs most relevant to assessment of polar processing diagnostics. Section 2.2 describes the diagnostics we calculate and the methods used to analyze them. Our results are presented in Section 3, comprising temperature (Section 3.1), vortex (Section 3.2) and derived (Section 3.3) diagnostics. Section 4 gives a summary and conclusions.

2 Data and Methods

2.1 Reanalysis Datasets

Fujiwara et al. (2017) provide detailed descriptions of the models, assimilation systems, and data inputs for the reanalyses used here in their overview paper on S-RIP. We compare the ~~five~~ four most recent high-resolution “full-input”~~reanalysis climatologies~~ reanalyses for all winters with data available since the beginning of the "satellite era" in 1979. All of our analyses are done using daily ~~12 UT~~ 12:00UT fields from each reanalysis dataset (we have tested the sensitivity of our analyses to using 00:00UT data and have found we get virtually identical results). Because of the importance of resolution, especially ~~the vertical grid~~in the vertical dimension, in representing the polar lower stratosphere and threshold processes in general (see, e.g., Manney et al., 2017), ~~except where unavailable (e.g., PV for MERRA),~~ we start our analyses from reanalysis data on the native model levels and at or (in the case of spectral models) near the native horizontal resolution.

2.1.1 ~~MERRA and MERRA-2~~

~~The National Aeronautics and Space Administration (NASA) GMAO's MERRA (Rienecker et al., 2011) dataset is a global reanalysis covering 1979 through 2015. It is based on the GEOS (Goddard Earth Observing System) version 5.2.0 assimilation system, which uses 3D-Var assimilation with Incremental Analysis Update (IAU) (Bloom et al., 1996) to constrain the analyses. The model uses a 0.5° × 0.667° latitude/longitude grid with 72 hybrid sigma-pressure levels, with about ~1.2km grid spacing in the lower stratosphere. PV data from MERRA were not archived on the model levels and grid, and thus we use the PV provided on a 1.25° × 1.25° latitude/longitude grid on 42 pressure levels.~~

~~MERRA-2 (Gelaro et al., 2017) uses a similar model and assimilation system to MERRA, with updates also described by Bosilovich et al. (2015) , Molod et al. (2015) , and Takacs et al. (2016) .Changes between MERRA and MERRA-2 that may significantly affect representation of the polar lower stratosphere include addition of new observation types in MERRA-2 (see Section 2.1.5, Figure 1 and Fujiwara et al. (2017)); a different treatment of conventional temperature data in MERRA-2; and assimilation of data in MERRA-2 using upgraded background error statistics, which control the magnitude and spatial extent of the impact of observations on the assimilated product. In addition, the effects of using a more uniform horizontal grid in MERRA-2 may be important. The impact of using the "cubed-sphere" grid in MERRA-2 is particularly seen in improved PV fields near the poles, especially in situations (common in the winter stratosphere) with strong cross-polar flow (e.g., see Figure 20 of Gelaro et al., 2017) . Different~~ Please see Table 1 for a list of relevant acronyms that we use below to describe the instruments, radiative transfer models~~are used for the assimilation of Stratospheric Sounding Unit (SSU) data in MERRA and MERRA-2. MERRA used GLATOVS (Goddard Laboratory for Atmospheres TOVS, Susskind et al., 1983) for SSU assimilation, while MERRA-2 uses the CRTM (Community Radiative Transfer Model, Han et al., 2006; Chen et al., 2008) for SSU and for all other radiance data. In particular, the CRTM takes into account long-term changes in the height of the SSU's weighting functions resulting from gradual reduction of pressure in the instruments' COcells (as described by, e.g., Kobayashi et al., 2009) GLATOVS did not.For the other radiances, MERRA also used CRTM, but a much older version than that used in MERRA-2. Gelaro et al. (2017) and references therein give details of these and other changes.~~

~~The MERRA-2 data products are described by Bosilovich et al. (2016) . All MERRA-2 data products used here are on model levels (the same vertical grid as for MERRA) and a 0.5° × 0.625° latitude/longitude grid. Data from MERRA-2 from its spin-up year, 1979, are not in the public MERRA-2 record, but we use data from late 1979 to start the analysis with the NH 1979/1980 winter. We use the MERRA-2 "Assimilated" (ASM) data collection (Global Modeling and Assimilation Office (GMAO), 2015) here, as recommended by GMAO, particularly for studies that require consistency between mass and wind fields (see, e.g., https://gmao.gsfc.nasa For MERRA, however, the ASM fields are not available on the model grid, but only at degraded horizontal and vertical resolution; we thus use the MERRA ANA collection for most variables here (see also Section 2.2.1). Differences between ANA and ASM fields are small, but can be non-negligible (e.g., Manney et al., 2017)~~ , etc., that are used by the reanalyses.

2.1.1 CFSR/CFSv2

NCEP-CFSR/CFSv2 ~~(hereinafter CFSR) (Saha et al., 2010)~~ is a global reanalysis ~~covering the period from~~ wherein CFSR covers 1979 ~~to the present~~ through 2010 and CFSv2 2011 through the present (Saha et al., 2010, 2014) . The data are produced using a coupled ocean-atmosphere model and 3D-Var assimilation. ~~CRTM is used~~ CFSR/CFSv2 uses the CRTM for satellite radiance assimilation. The model resolution is T382L64~~;~~, but the data used here are on a 0.5° × 0.5° horizontal grid on the model levels (available through 2015); vertical grid spacing in the lower stratosphere ranges from about 0.8 km near 100 hPa to 1.3 km near 10hPa. CFSR did make an undocumented update to their assimilation scheme in 2010 (Long et al., 2017). Furthermore, in the transition from CFSR to CFSv2 in 2011, the resolution, forecast model, and assimilation scheme were all upgraded; CFSv2 is, however, intended as a continuation of CFSR and can be treated as such for most purposes (Saha et al., 2014; Fujiwara et al., 2017; Long et al., 2017); we thus treat these as a single reanalysis in this paper.

2.1.2 ERA-Interim

ERA-Interim (see Dee et al., 2011) is another global reanalysis that covers the period from 1979 to the present. The data are produced using 4D-Var assimilation with a T255L60 spectral model. ERA-I uses the ~~RTTOV (Radiative Transfer for TOVS)~~ version 7 RTTOV radiative transfer model ~~, version 7~~ for radiance assimilation. Here we use ERA-Interim data on a 0.75° × 0.75° latitude/longitude grid (near the resolution of the model's Gaussian grid) on the 60 model levels. The spacing of the model levels in the lower stratosphere is about 1.2 to 1.4 km.

2.1.3 JRA-55

JRA-55 (Ebita et al., 2011; Kobayashi et al., 2015) is a global reanalysis that covers the period from 1958 to the present, and is produced using ~~a~~ 4D-Var assimilation. The data from the JRA-55 T319L60 spectral model are provided on an approximately 0.56° Gaussian grid corresponding to that spectral resolution. JRA-55 uses RTTOV version 9.3 for satellite radiance assimilation. The JRA-55 fields on the model vertical levels have a vertical resolution of ∼1.2 to 1.4 km in the lower stratosphere.

2.1.4 MERRA-2

MERRA-2 (Gelaro et al., 2017) is a global reanalysis produced by the National Aeronautics and Space Administration Global Modeling and Assimilation Office (NASA GMAO) covering 1980 to the present. It is based on the Goddard Earth Observing System (GEOS) 5.12.4 assimilation system, which uses 3D-Var assimilation with Incremental Analysis Update (IAU) (Bloom et al., 1996) t constrain the analyses. MERRA-2 is intended to be a replacement for its predecessor, MERRA (Rienecker et al., 2011) , as it includes many updates over MERRA (see, e.g. Bosilovich et al., 2015; Molod et al., 2015; Takacs et al., 2016) . Changes between MERRA and MERRA-2 that may significantly affect representation of the lower stratosphere include the addition of new observation types in MERRA-2 (see Section 2.1.5, Figure 1 and Fujiwara et al. (2017)); an updated radiative transfer model for radiance assimilation; a different treatment of conventional temperature data; and assimilation of data that uses upgraded background error statistics, which control the magnitude and spatial extent of the impact of observations on the assimilated product.

The MERRA-2 data products are described by Bosilovich et al. (2016) . All MERRA-2 data products used here are on 72 hybrid sigma-pressure levels that have about ∼1.2 km grid spacing in the lower stratosphere, and a 0.5° × 0.625° latitude/longitude grid. Data from MERRA-2 from its spin-up year, 1979, are not in the public MERRA-2 record, but we use data from late 1979 to start the analysis with the NH 1979/1980 winter. We use the MERRA-2 “Assimilated” (ASM) data collection (Global Modeling and Assimilation Office (GMAO), 2015) here, as recommended by GMAO, particularly for studies that require consistency between mass and wind fields (see, e.g., Global Modeling and Assimilation Office (GMAO), 2017; Fujiwara et al., 2017) .

2.1.5 Timeline of Satellite Data Inputs to DAS

Operational satellite observations are the primary data constraints on reanalyses at stratospheric levels. Additional constraints on temperature are provided by radiosonde and other conventional observations (Fujiwara et al., 2017) . While conventional data are important in the lower stratosphere, especially at midlatitudes, their coverage is sparse in the NH polar region and very poor in the SH so that these regions are mainly constrained by satellite radiance measurements. Figure 1 shows these satellite inputs (see Table 1 for ~~a list of satellite~~ the definitions of the acronyms) for each of the reanalyses used here, shown as stacked timelines to facilitate comparison of changes in data inputs between reanalyses. Up through about 1994, all of the reanalyses relied primarily on the TOVS instruments (SSU, MSU, VTPR & HIRS), and (excepting CFSR/CFSv2) SSM/I & SSMIS. Between 1995 and 2002, there are several changes in the data inputs, and the inputs begin to vary more among the reanalyses. ~~MERRA and MERRA-2 have much the same inputs for most of the period, but for~~ In recent years MERRA-2 assimilates Meteosat, IASI, CrIS, ATMS, and GPS-RO observations. IASI, CrIS, and ATMS are not assimilated in any of the other reanalyses. A change with a large impact on stratospheric temperatures overall is the transition from TOVS (with MSU and SSU) to ATOVS (with AMSU-A and AMSU-B/MHS); Figure 1 shows that this transition is handled/timed differently among the reanalyses: For instance, although all reanalyses introduce AMSU-A at just about the same time, JRA-55 stops assimilating the TOVS instruments’ data by 2000, whereas the others continue until 2006 (except for an immediate cut off of SSU by 1999 in CFSR/CFSv2). Long et al. (2017) showed that this transition had a profound impact on the differences in zonal mean temperatures among the reanalyses, with a shift toward much better agreement after the transition.

2.2 Methods

The methods and diagnostics used herein are largely the same as those used by Lawrence et al. (2015). In the subsections to follow, we describe the reanalysis fields we use, and how we prepare them and derive the polar processing diagnostics from them. We also provide some information on additional analysis techniques that we use to help interpret our results.

2.2.1 Preparation of Meteorological Fields

The two meteorological fields necessary to derive most of the diagnostics used herein are temperatures and potential vorticity (PV) on isentropic surfaces. In the results we present later, we show temperature diagnostics on pressure surfaces, and vortex diagnostics on isentropic surfaces; as will be discussed, we also calculate and use some temperature diagnostics on isentropic levels.

Of the ~~5~~ four reanalyses described in Section 2.1, only MERRA-2 provides potential vorticity on their model levels. ~~MERRA provides PV calculated within the assimilation system, but interpolated to 42 pressure levels on a reduced horizontal resolution grid, which we interpolate to the same grid (the native model grid) on which we use the MERRA temperatures.~~ ~~CFSR~~CFSR/CFSv2, ERA-Interim and JRA-55 ~~only provide PV~~ have isentropic PV available, but these products are only provided on a sparse set of isentropic levels, with very few common levels between ~~them~~the reanalyses. ERA-Interim provides absolute vorticity on model levels, and CFSR/CFSv2 provides relative vorticity; thus to get the vertically-resolved PV fields that we need, we derive PV for these reanalyses using their provided vorticity, temperature, and pressure fields on the model levels. In the case of JRA-55, we use the zonal and meridional wind components to first calculate relative vorticity, which we then use in combination with temperature and pressure to calculate PV on the model levels. While a thorough evaluation of biases that may arise from using different types of PV calculations for different reanalyses would be valuable, it is beyond the scope of this paper. We think, however, that data users will most likely use the most direct calculation to get from the provided fields to the model-level PV (as we have), and thus the fields we are comparing are those that are most likely to be used in practice. When calculating polar processing diagnostics, we scale the PV fields into vorticity units as in Dunkerton and Delisi (1986) by dividing the PV by a standard value of static stability calculated from assuming a vertical temperature gradient of $1\,\mathrm{K\,km^{-1}}$ and a pressure of 54 hPa on the 500 K isentropic level. We use this scaling so that the scaled PV (sPV) values are of the same order of magnitude at different levels throughout the stratosphere.

Since the reanalyses are all on different model levels, we use the reanalyses' temperature and pressure fields to vertically interpolate their temperature and PV fields to a common set of fixed pressure and isentropic surfaces. We use a standard set of pressure and isentropic levels that have been used for several NASA satellite instrument datasets (see, e.g., https://cdn.earthdata.nasa.gov/conduit/upload/4849/ESDS-RFC-009.pdf) including the Aura Microwave Limb Sounder (Livesey et al., 2015); these are 12 levels per decade in pressure, and their climatologically corresponding isentropic surfaces. For polar processing diagnostics, we limit our focus to ~~the thirteen levels between roughly~~ fourteen levels with pressures between approximately 120 ~~–~~ and 10hPa~~, or 390 – 850~~; these pressure levels and their corresponding isentropic surfaces are discussed in Appendix A.

2.2.2 Temperature and Vortex Diagnostics

The depletion of ozone in the lower stratosphere follows from a complex chain of processes that are highly dependent on meteorological conditions (see, e.g., Solomon, 1999; WMO, 2014, for reviews). The activation of chlorine requires the presence of polar stratospheric clouds (PSCs), which form when temperatures are sufficiently low, and grow when temperatures stay low for sufficiently long periods of time (Hanson and Mauersberger, 1988; Solomon, 1999; WMO, 2014, and references therein). The catalytic ozone destruction cycles involving both chlorine and bromine further require sunlight, which is usually provided later in winter and spring when sunlight returns to the high latitude polar regions. These chemical processes also require isolation from lower latitude air, which is provided by the stratospheric polar vortex; the edge of the polar vortex acts as a barrier preventing transport and mixing, and thus the vortex acts as a containment vessel for the polar air where these processes take place (e.g., Schoeberl and Hartmann, 1991; Schoeberl et al., 1992). We thus examine polar processing diagnostics that primarily focus on lower stratospheric temperatures and the state of the polar vortex to assess the meteorological conditions conducive to ozone depletion. Unless specified otherwise, we focus on the months from December through March (DJFM) for the Northern Hemisphere (NH), and May through October (MJJASO) for the Southern Hemisphere (SH); these time periods cover roughly the full period during which polar processing takes place for most polar winters (see Section 3.1).

The temperature diagnostics that we use include minimum temperatures (T_{min}) poleward of ±40° latitude, and the areas of temperatures (poleward of ±30° latitude) below PSC formation thresholds (A_{PSC}, or Area $T \leq T_{PSC}$, where the subscript 'PSC' may be replaced by 'NAT' or 'ice' to denote the specific type of PSC). For A_{PSC}, we specifically use the formation temperatures for solid nitric acid trihydrate (NAT, Hanson and Mauersberger, 1988) and ice particles on pressure levels, which we define using climatological profiles of HNO_3 and H_2O mixing ratios (see Appendix A and Table A1 for values). As a rule of thumb, the NAT threshold between 120 and 10 hPa ranges from roughly 198 – 187 K respectively, and the ice threshold tends to be between 6 – 8 K below the NAT threshold. We stress that these thresholds are approximations, but they are convenient as proxies for PSC formation and chlorine activation. While most of the results we show herein for the temperature diagnostics are on pressure levels, we also calculate them on isentropic surfaces; for the diagnostics involving PSC thresholds, we assign the PSC thresholds on pressure levels to the isentropic surfaces that are roughly co-located (e.g., 520 K corresponds to 46.1 hPa). This is an additional approximation, but it allows us to keep the intercomparisons simple without having to calculate daily varying PSC thresholds for pressures/temperatures on isentropic levels, or pre-computing climatological PSC threshold values from the reanalyses' fields. (Please see Appendix A and Table A2 for the pressure surfaces used, their corresponding isentropic surfaces, and the PSC threshold temperatures.) To mitigate issues with these approximations and test sensitivity to the thresholds, we also compute A_{PSC} with ±1 K offsets to the PSC formation temperatures.

The vortex diagnostics that we use include maximum gradients in sPV as a function of equivalent latitude (maximum sPV gradients, or MPVG), which assess the strength of the vortex edge (e.g., Manney et al., 1994, 2011; Lawrence et al., 2015), and the area of the vortex exposed to sunlight (sunlit vortex area, or SVA). To calculate MPVG, we bin sPV as a function of equivalent latitude (EqL, the latitude that would enclose the same area between it and the pole as a given PV contour; Butchart and Remsberg, 1986), numerically differentiate, and catalog the maximum value between ±30 and ±80° EqL. We use ±30°as a lower limit because relatively large PV gradients can be found in the tropics, which can dominate early/late in the season when the vortex is forming/decaying; we use ±80°as an upper limit because the small areas represented by points poleward of ±80°can vary much more dramatically than the lower EqLs, sometimes producing large gradients (e.g., Nash et al., 1996) that are not indicative of the vortex edge.

To calculate SVA, we calculate the area of the vortex that extends equatorward of the daily polar night latitude at 12:00UT. ~~We use the vertical profile of~~ The area of the vortex is determined by defining constant contours of sPV as the vortex ~~edge values defined by Lawrence and Manney (2018) rounded to the nearest 1.0~~ edge over all years; herein we determined these vortex edges individually for each reanalysis from climatological seasonal averages of their maximum PV gradients (as determined above) from the extended periods of November through April for the NH, and April through November for the SH. We use periods that are longer than the DJFM and MJJASO periods we use for the intercomparisons because they help to include the formation and breakdown of the vortex. While constant vortex edges are a simplification, the ones we use here are defined for each reanalysis individually, and thus they inherently fold in any systematic differences that the reanalysis sPV fields may have. Furthermore, more common definitions of the vortex edge that provide daily varying values can be prone to give spurious oscillations from day to day that could contaminate intercomparisons (e.g. Manney et al., 2007; Lawrence and Manney, 2018) . Figure 2 shows the NH and SH profiles of vortex edges used for each reanalysis; it shows that the values obtained for each reanalysis and hemisphere are generally consistent, and that the largest differences between reanalyses below 850 K are around 0.2 × 10~~$^{-5}$~~$^{-4}$ s^{-1}~~to specify the vortex edge in both the NHand SH (in the SH, the values are multiplied by -1). This vortex edge profile was defined from data for the NH polar vortex , but we have confirmed it is appropriate for the SH vortex, whose edge varies (in a climatological sense) similarly with height~~.

2.2.3 ~~Analysis Techniques~~

~~Rather than comparing the diagnostics derived from each of the reanalyses to an average across a subset of the reanalyses (for S-RIP, this is referred to as the "Reanalysis Ensemble Mean"), we opt for comparing , , ,~~

2.2.3 Derived Diagnostics

Here we describe some additional diagnostics that we examine later in the paper that are derived from the raw diagnostics we calculate (primarily those described above).

The winter mean volume of lower stratospheric air with temperatures below T_{PSC} (V_{PSC}) is a widely used diagnostic of polar processing potential. It is often expressed as a fraction of the vortex volume (V_{PSC}/V_{Vort}) to provide a measure that is independent of the substantial interannual and interhemispheric variations in vortex size (e.g., Rex et al., 2004, 2006; Tilmes et al., 2006; M (Again, we replace the subscript 'PSC' by 'NAT' or 'ice' to denote specific PSC types.) Hence, V_{PSC}/V_{Vort} represents the approximate fraction of the vortex (for a specified altitude range) in which temperatures are low enough for the formation of PSCs. Here, we calculate V_{PSC} and the volume of the vortex using A_{PSC} and the area of the vortex on isentropic levels between 390 and ~~directly with . We do this for a few reasons: first, 4 – 5 reanalyses is a relatively small sample that can make the ensemble average sensitive to outliers. The second and related reason we compare with directly is that comparing with an average across the reanalyses obscures the actual scale of differences among the individual datasets since the mean of~~ 550 K. A_{PSC} is calculated as described above in Section 2.2.2; the area of the vortex is calculated similarly to SVA, but instead by finding the total area within the PV contours representing the vortex edge. To get volumes, we assume each isentropic level is nominally representative of the volume of air midway between each level; for example, the ~~reanalyses will "centralize" the~~

~~diagnostics. Finally, it is unclear to us whether it is appropriate to include both or only one of or in the average; while is an improvement over with numerous differences in the data ingested, they do use similar models, assimilation schemes, etc. Thus, for the results herein, we will primarily show climatologies and reanalysis differences from~~ 410 K level comes after 390 K and before 430 K, so 410 K is assumed to be representative of the altitude “width” between 400 and 420 K. The altitude widths of these nominal levels are determined using the Knox (1998) approximation; for the levels from 390 to 550 K, the Knox approximation gives a mean altitude differential between levels of 1.13 km with a minimum of 0.98 km, and a maximum of 1.30 km. These altitude differentials are then multiplied by the area diagnostics on each isentropic level (which are converted to km^2), and summed over the vertical range to get volumes. The volume fraction is then V_{PSC}/V_{Vort}. In the results we show later on, we specifically show winter mean V_{PSC}/V_{Vort}; these winter means are taken over DJFM for the NH, and MJJASO for the SH.

The SH vortex breakup is of considerable concern because it results in the dispersal of ozone-depleted vortex air over mid-latitudes (e.g., Ajtić et al., 2003, 2004; Manney et al., 2005c; Pazmino et al., 2005; WMO, 2007) . While ozone depletion in the Arctic has not yet been large enough for this to be an ongoing concern, vortex evolution during the 2011 Arctic vortex breakup led to significant areas of ozone depleted air over populated regions associated with increased surface UV (e.g., Manney et al., 2011; Bernhard et al., 2012) . To examine the variability and representation in reanalyses of the vortex decay in the lower to middle stratosphere, we examine approximate vortex decay dates, which we derive using the vortex area diagnostic on isentropic levels from 460 to 850 K. Here we calculate vortex area with $+0.1 \times 10^{-4}$ s^{-1} sPV offsets to the vortex edges shown in Figure 2. To accomplish this, we examine NH vortex area between 1 Dec and 1 Jun, and SH vortex area between 1 May and 1 Mar; we have defined the decay date as the last day before which the vortex area is above 1% of a hemisphere continuously for 30 days. We choose 1% of a hemisphere as the limit because this threshold is only climatologically met at all levels at the beginning and end of the seasons when the vortex is forming or breaking down, which guarantees that any time the vortex is that small, it is either significantly disturbed or in the process of decaying. The 30 day limit was chosen to help guarantee that the vortex was sufficiently coherent beforehand. Finally, we use vortex edges with the positive sPV offset mentioned above to help remove the influence of small vortex fragments that can be present at the end of the season, which in some cases can add up to areas larger than 1% of a hemisphere and lead to marginal scenarios that can skew the decay dates. The results we show herein are not highly sensitive to changing the area threshold or using vortex area with/without the sPV offset; except in some marginal cases that we discuss later on, adjusting the area threshold between 1 and 4% only modifies less than 10% of the cases (i.e., ~~reanalysis minus) . Henceforth, we will colloquially refer to the group of , , , and as “the reanalyses ”, but this should not be confused as us making a value judgment of~~ different years and levels) in all the reanalyses by more than 20 days in the NH, and more than 10 days in the SH.

2.2.4 Analysis Techniques

For most of the results shown herein, we compare the diagnostics derived from each of the reanalyses to an average across all of the reanalyses, which is referred to as the “~~best” or “standard” reanalysis for other reanalyses to be compared against~~Reanalysis

Ensemble Mean" (the REM). In Sections 3.1 and 3.2, the comparisons primarily take the form of reanalysis differences from the REM (i.e., reanalysis minus REM).

~~Part of our analysis uses~~ Our analysis also includes a statistical significance test to determine whether the average differences between the reanalyses and ~~MERRA-2~~ the REM are statistically different from zero over a winter season. To accomplish this, we use a non-parametric bootstrap resampling technique that is useful for time series datasets called the stationary bootstrap (Politis and Romano, 1994). Bootstrapping methods for time series have generally relied on resampling blocks of consecutive observations to construct many artificial time series so that accuracy estimates can be made for sample statistics/estimators (e.g., Lahiri, 2003, and references therein). Rather than resampling random fixed-size blocks (which may or may not overlap) ~~of a fixed size~~ to construct artificial time series, the stationary bootstrap constructs ~~pseudo~~ artificial time series by resampling ~~blocks of random size~~ random blocks with random sizes determined from a geometric distribution with specified mean. Herein, we bootstrap the time series of differences from the reanalyses and the REM; we treat the ~~differences from~~ difference time series for each reanalysis, ~~on each vertical level, and each year individually . In all cases, we use the stationary bootstrap~~ diagnostic, and year individually while the vertical levels are resampled together. In nearly all cases (see the NH A_{NAT} comparisons in Section 3.1 for the one exception), we perform stationary resampling with a specified geometric distribution mean of ~~5~~ 10 (i.e., the expected block length is ~~5~~ 10 days), and resample all the time series of differences 2×10^5 times. We note that the results shown herein are not sensitive to the choice of the expected block length; we repeated our bootstrapping analysis for different expected block lengths between 5 and 15 days, and in all cases the results were nearly identical. Ultimately we chose 10 days as a happy medium based on examinations of the decorrelation time scales of some of the difference time series. We then use the bootstrap percentile method to construct 99% confidence intervals (CIs) of the average differences; the percentile method is known to have issues in cases with small sample sizes, but since we use a more strict 99% CI and our time series are longer than 120 days, we expect our estimates are robust (see discussion in DiCiccio and Efron, 1996, and references therein). When these 99% CIs do not contain zero, we consider the average differences for the reanalysis minus the REM (for a specific level and year) to be indicative of persistent positive or negative differences.

3 Results

In the next two subsections, we show comparisons of temperature and vortex diagnostics as yearly time series of average differences and standard deviations calculated over the polar processing periods in each hemisphere (DJFM for the NH, MJJASO for the SH). We use these averages and standard deviations ~~to evaluate the agreement between the reanalyses. To demonstrate what we mean by agreement, in Figure ?? we have plotted yearly time series of the average differences and standard deviations of differences in a theoretical diagnostic calculated by subtracting one comparison reanalysis from two theoretical reanalyses (cyan and magenta lines). The magenta reanalysis tends to have smaller average differences from the comparison reanalysis throughout the period than the cyan reanalysis, but the cyan reanalysis tends to have smaller standard deviations. Even when the magenta reanalysis has relatively small average differences between 1995 and 2000, it has large standard deviations, indicating that the spread around the average is quite large during this time. In contrast, for the same~~

~~period, the cyan reanalysis has relatively small standard deviations, but large negative average differences, indicating that the diagnostic calculated from the cyan reanalysis tends to be systematically higher than the comparison reanalysis. Later on, the average differences and standard deviations for both reanalyses approach zero, indicating a convergence towards better agreement with the comparison reanalysis, but also better~~ alongside the bootstrapping analysis to evaluate the agreement between the ~~cyan and magenta reanalysesthemselves~~reanalyses.

~~Since we examine averages and standard deviations of differences across multiple years and vertical levels, we illustrate in Figure ?? how we display the reanalysis differences in Sections 3.1 and 3.2. For each year, level, and reanalysis, the diagnostics calculated from (minimum temperatures in this example; top panel) are subtracted from those calculated from the other reanalyses (in the example; bottom left panel). The averages and standard deviations are then found over the full polar processing period bounded by the vertical black lines) for each vertical level. These values are then plotted as individual pixels in a column (bottom right panel) summarizing the differences from a single year.~~

3.1 Temperature Diagnostics

Figure 3 shows the climatological values of minimum temperatures from ~~MERRA-2~~the REM. The well known difference in stratospheric temperatures between NH and SH (e.g., Andrews, 1989) is seen clearly, with the climatological period with temperatures below the NAT PSC threshold spanning approximately December through mid-February in the NH and mid-May through early October in the SH. The lowest temperatures are centered near 20 hPa at about the time of the solstice in the NH, and near 25 hPa approximately a month after the solstice in the SH. NH winter temperatures are lowest earlier in the season because of the prevalence of sudden stratospheric warmings (SSWs) in January and February in that hemisphere.

Figure 4 shows ~~the yearly time series of~~ “pixel plots” of the winter mean differences in ~~minimum temperatures for the Antarctic, as "pixel " plots constructed as described above.~~ SH minimum temperatures from the REM (left column), and the standard deviations of the differences (right column) for each of the reanalyses. We use similar pixel plots herein for the other diagnostics and hemispheres discussed in Section 2.2.2. In these plots, each pixel represents a winter mean difference (i.e., reanalysis minus REM averaged over a winter period) or a standard deviation of the differences (i.e., the standard deviation of the reanalysis minus REM over the designated winter period) for a single year and vertical level.

The most striking feature shown in Figure 4 is an overall improvement in the agreement ~~after 1998~~ around the turn of the century particularly evident in MERRA-2 after 1998. This transition is also apparent in ERA-I, occurring between 1999 and 2001. In earlier years ERA-I and MERRA-2 bracket the ensemble with differences up to ~~1999, with some differences having magnitudes up to about 6~~±3 K~~in the earlier years , as opposed to near 1~~, which in later years drop to near 0.5 K~~in later years . A corresponding decrease is seen~~. The SH CFSR/CFSv2 and JRA-55 minimum temperatures tend to reside between those of MERRA-2 and ERA-I and are generally close to the REM. In particular, the JRA-55 differences are marked as not statistically significant for many levels and years throughout the reanalysis period. The improvements after 1998 are largest at higher levels (where the differences and standard deviations are themselves largest), becoming less prominent, and less sudden, below about 50 hPa. MERRA-2 shows a change in sign of the differences in the upper levels (∼20–10 hPa). The overall convergence of the reanalyses after 1998–1999 is also seen as pronounced discontinuities in the standard deviations of the differences from the

REM for ERA-I, JRA-55 and MERRA-2, with values ~~up to over 3~~frequently over 2 K before 1999 ~~, and typically near 1~~typically decreasing to below ∼0.8 K thereafter. ~~This time~~ The improvement is less evident in CFSR/CFSv2 with standard deviations greater than 1 K seen throughout the reanalysis period, particularly at pressures lower than ∼ 30 hPa. The 1998–1999 mark corresponds to the transition from assimilating TOVS to ATOVS radiances in all four reanalyses. In addition, in late 2002 ~~, all of the reanalyses except JRA-55~~ MERRA-2 began assimilating the hyperspectral AIRS radiances, vastly increasing the number of ~~observations~~ data used in the Antarctic. ERA-I and CFSR/CFSv2 started ingesting AIRS data in 2004. Measurements of atmospheric thermal emissions in the 15-μm CO_2 absorption continuum provided by this sensor are strongly sensitive to stratospheric temperature variations as demonstrated, for example, by Hoffmann and Alexander (2009) . In addition, all the reanalyses considered here assimilate data from GPS-RO instruments starting in 2001 in ERA-I and CFSR/CFSv2, in 2004 in MERRA-2 and in 2006 in JRA-55. The GPS-RO data also affect stratospheric temperatures indirectly by anchoring bias correction of radiance observations (Fujiwara et al., 2017) .

During years from roughly 1993 through 1998, larger differences and standard deviations are seen above about 30 hPa in ~~ERA-I, and to a lesser degree in the other reanalyses. While the main source of stratospheric information for all the reanalyses in this time period is the SSU and MSU instruments, there are several differences in how these data are assimilated: The different radiative transfer models used for satellite radiance assimilation handle inter-satellite drifts due to SSU COcell pressure leaks differently; MERRA and MERRA-2 also handle these differently because of the change in radiative transfer model used. There are also differences in which SSU channels a bias correction is applied to (MERRA-2 does not bias correct Channel 3, but JRA-55 and ERA-I do). It is even more difficult to speculate about changes in CFSR, since it has multiple discontinuities and biases related to stitching together execution streams and applying a bias correction in a model with a warm bias (Long et al., 2017) . Thus, while we cannot pin down particular changes that are associated with this increase in variance, there are numerous differences that could contribute to this behavior.~~

ERA-I

and MERRA-2. These differences are positive in the former and negative in the latter reanalysis, leading to a partial cancellation in the REM.

~~The improvements after 1998 are largest at higher levels (where the differences and standard deviations are themselves largest), becoming less prominent, and less sudden, below about~~ Between 1986 and 2001 ERA-I exhibits a layered structure of differences from the REM: positive at pressures greater than ∼50 hPa ~~. Before 1999, CFSR, ERA-I, and JRA-55 all show negative differences between about 70 and~~ and less than ∼30 hPa ~~sandwiched between positive differences above and below that.~~ and negative in between. A similar structure but with the signs reversed is seen in MERRA-2 where it extends back to 1980 and ends sharply in 1998. Investigations in progress (Long et al., in preparation) show that both ~~MERRA-2 and ERA-I~~ MERRA-2 and ERA-I temperatures in the SH polar stratosphere have oscillations of up to about 3 K ~~, which~~ that are in opposite directions, leading to the ~~layered~~ structure of the differences seen here. ~~The other reanalyses do not show vertical oscillations, but the oscillations in MERRA-2, of course, show up in the differences. (Note that the absence of oscillations in the other reanalyses does n~~ (Note that the absence of oscillations in the other reanalyses does not imply better agreement with sondes; Long et al., in preparation) . After 2000 both reanalyses show slightly positive (and, in the case of MERRA-2, largely statistically insignificant) differences from the REM at most pressure levels. CFSR/CFSv2 shows mostly positive differences between 1979 and 1986; afterward the differences are primarily slightly negative at most of the pressure levels shown.

While the main source of stratospheric information for all the reanalyses before 1998 is the SSU and MSU instruments, different reanalyses use different radiative transfer models to assimilate them and apply bias correction differently (Wright et al., in preparat It is particularly difficult to speculate about changes in CFSR/CFSv2, since it has multiple discontinuities and biases related to stitching together execution streams and applying a bias correction in a model with a warm bias (Long et al., 2017) . Thus, while we cannot pin down particular changes that are associated with the differences among the reanalyses prior to the introduction of ATOVS data, there are numerous factors that could contribute to this behavior.

~~Average differences between MERRA-2 and ERA-I, JRA-55, and MERRA are not statistically significant for many regions and years after 1999, especially between about 50 and 20hPa. In comparison with the other reanalyses, CFSR continues to show larger standard deviations and more significant average differences from MERRA-2 after 1999. CFSR and JRA-55 show a change in sign of the differences in the upper levels(~20–10hPa), while ERA-I shows a similar change in sign, but only above about 15hPa. After 1999, the MERRA − MERRA-2 differences also change signs from negative (positive) to positive (negative) at pressures less (greater) than roughly 40hPa; however, their differences during this time period are typically within 1K. Despite substantial changes in the model and assimilation systems, MERRA shows much closer agreement with MERRA-2 before 1998 than do the other reanalyses, but still shows a sudden decrease in the differences at that time. This suggests that the improved resolution from the ATOVS instruments and the increase in the number of observations are major factors in the improved agreement among all reanalyses, but that the differences in the models and data handling (which are smaller between MERRA and MERRA-2 than between the other reanalyses and MERRA-2) are also an important factor.~~

~~The year 1984 stands out, especially in JRA-55, as having much lower differences and larger standard deviations in the higher levels than during the surrounding years. Inspection of the daily differences for individual years (not shown) indicates a period in July, August, and September in that winter with negative differences instead of large positive ones in CFSR, JRA-55, and MERRA, and a period in June and July in ERA-I with much smaller positive differences than those in other years, at the highest levels. While in 1983 and 1985, MERRA-2 assimilated radiances from two SSU instruments and several channels from those instruments, in 1984 MERRA-2 assimilated data from only one (NOAA-7) SSU instrument, and channel 2 (which peaks above, but has considerable influence below, 10hPa) on that instrument was off. Furthermore, there was a change in how MERRA-2 assimilated MSU (whose highest peaking channel samples the lower stratosphere) radiances at about this time. While not conclusive, these changes could have contributed to the anomaly in the MERRA-2 differences from the other reanalyses in 1984.~~

Average differences in minimum temperatures in the NH (Figure 5) show more complicated patterns of changes over the years than those seen in the SH. The differences are much smaller throughout the ~~36-year~~ 38-year period, with maximum ~~differences near 2~~absolute differences near 1.5 K at the highest levels shown ~~,~~ (mainly in the period from about 1994 to 2004), and more frequent ~~times~~years/~~regions throughout the period studied when~~ levels where the average differences are not ~~significant. The~~ statistically significant. From roughly 10–25 hPa, the standard deviations do decrease ~~somewhat~~ from above ~1 K to less than ~0.75 K after around 1999, though (as was the case in the SH) they remain larger in ~~CFSR~~ CFSR/CFSv2 than in the other reanalyses. ~~There~~ While there are indications of ~~sudden~~ changes around 1999, ~~but~~ particularly in ERA-I and MERRA-2 they are less abrupt and of smaller magnitudes than those in the SH, and it is ~~often less~~ not as clear that there is a uniform trend towards better agreement. ~~Prior to about 1999, ERA-I, JRA-55, and CFSR show a similar pattern to that~~ As in the SH~~of positive differences at the lowest and highest levels surrounding negative differences between those layers; CFSR and JRA-55 show more regions of negative than positive differences after 1999. ERA-I shows a decrease in the regions with significant average differences after about 1999, but , while the patterns of such regions change for the other reanalyses, a clear decrease in them is not evident. As in the SH , the differences between MERRA-2 and MERRA are smaller than those between MERRA-2 and~~ , the CFSR/CFSv2 differences are primarily positive before 1987 and negative afterwards. ERA-I shows mostly positive differences except near 10 hPa. The opposite is true for JRA-55, except for the period between 1998 and 2006 when the differences are near zero, but slightly positive (and many of them not statistically significant), at pressures greater than about 30 hPa. Similar to the SH case, MERRA-2 exhibits a layered structure of differences prior to 1998: positive between roughly 60 and 30 hPa and negative outside of this layer. After 1998 the ~~other reanalyses in the period before about 1998.~~ MERRA-2 differences are mainly positive, except at pressures greater than about 50 hPa where the differences gradually change from negative to positive between 2005 and 2010, except at the lowest levels.

Figure 6 shows ~~MERRA-2~~ the REM climatological values of the area with temperatures below the NAT PSC threshold (A_{NAT}) for the NH and SH winter seasons. As expected, these echo the patterns of minimum temperatures seen in Figure 3, with the largest ~~values~~ areas in the NH in early January, and in the SH in middle to late July. The great variability in the NH (see the grey envelopes in the line plots) results in the largest values being well above the climatological average, about 7–8% of a hemisphere, but still much lower than the largest average values in the SH of over ~~12~~10% of a hemisphere.

Note that comparing differences in NH A_{NAT} among the reanalyses is more difficult than doing so for the SH or for the other NH diagnostics. Because there is significant interannual variability in the onset, termination, and magnitudes of low temperatures in the NH (see both Figures 3 and 6), there are many NH winters with relatively few days having temperatures below T_{NAT}, and thus many days with NH A_{NAT} being zero. Thus, comparing differences among the reanalyses for the full DJFM time period can often be unfairly biased by the high occurrence of zeros, which artificially decreases the average differences and standard deviations. To allay this issue such that we fairly compare NH A_{NAT}, we modify our analysis procedure as follows: We use time series of the REM NH A_{NAT} from November through April on 30, 50, and 70 hPa to define approximate start and end dates for the periods having non-zero A_{NAT}. We use 30 hPa to define the onset dates (because A_{NAT} usually first becomes nonzero around this level; e.g., Figure 6a), and 50 hPa or 70 hPa to define the termination dates. More specifically, we define the onset dates for each year as the first day at 30 hPa having nonzero A_{NAT}, and the termination dates as the latest day chosen by either 50 or 70 hPa having nonzero A_{NAT}; both 50 and 70 hPa are used because termination most often happens latest around 70 hPa as seen in Figure 6a, but in some winters it happens later around 50 hPa. This process gives us individual "NAT seasons" between 1979/1980 and 2016/~~17 having~~ 2017; these have a median length of ~~87~~ 85 days, with the minimum and maximum number of days being ~~41~~ 40 and 126, respectively. We then use these truncated time series to define the average differences and standard deviations thereof. This modifies the bootstrapping procedure described in Section 2.2.4; we still perform 2×10^5 stationary bootstraps for each year, but because the lengths of the time series vary, we also vary the expected block size for each year by specifying them as the nearest integer to the cube root of the time series lengths plus a constant offset of +5 (which ranges from ~~3 to 5~~ 8 to 10 days for time series lengths between ~~41~~ 40 and 126 days). As was found for the regular bootstrapping procedure, using different expected blocklengths with offsets between 0 (3 to 5 days) and 10 (13 to 15 days) had very little effect on the statistical significance results.

Figure 7 shows A_{NAT} differences ~~for~~ from the REM for the SH winter seasons. ~~As was the case for the minimum temperatures, there is a large decrease in both the magnitude of the average differences and the standard deviations after 1998. The changing signs~~ There is a very apparent sudden decrease in the seasonal standard deviations of the differences ~~with height before 1999 are also consistent with those seen in the minimum temperatures, with positive differences between about 50 and 20~~at levels above ∼25 ~~hPa sandwiched between negative differences above and below for CFSR, ERA-I, and JRA-55. Furthermore, the increase in average differences in the upper levels between 1993 and 1998 is again apparent~~ hPa after 1998 similar to, but much more pronounced than in the case of minimum temperatures. The ~~average differences generally show more regions that are not significantly different from zero in the later years, though the patterns of this close agreement are rather different than those for the~~ 1998–1999 boundary is less obvious in the average differences for CFSR/CFSv2 and JRA-55, but is apparent in ERA-I and MERRA-2. By these metrics, all four reanalyses converge toward better agreement following the TOVS/ATOVS transition. The patterns of differences largely mirror (in an opposite sense) the patterns shown in Figure 4; that is, there tend to be positive/negative differences from the REM in A_{NAT} wherever there are negative/positive differences from the REM in minimum temperatures. ~~For this diagnostic, ERA-I continues to show significant average differences for most levels/regions throughout the period, though the magnitudes of~~ ERA-I and MERRA-2 display layered difference structures prior to 1998; these layers of positive and negative differences are separated by approximately the 30 and 70 hPa pressure levels. As in the case of the minimum temperatures, the layered structures are more persistent in MERRA-2, extending between 1980–1998, whereas the one in ERA-I becomes apparent after 1986. JRA-55 and CFSR/CFSv2 are more often closer to the REM in terms of both the mean differences and the standard deviations. For CFSR/CFSv2 at pressures greater than 20 hPa the differences are ~~much smaller than those in the earlier years. The differences that are seen between the patterns of T_{min} versus A_{NAT} average differences suggest some minor differences between reanalyses in the morphology of the fields (e.g., spatial patterns or gradients) beyond just overall temperature biases~~mostly negative (smaller A_{NAT}) prior to 1986, and mostly positive thereafter. No clear pattern is apparent for JRA-55, although after approximately 2005 each reanalysis generally has a uniform sign of the differences from the REM in the deep layer between 120 and 10 hPa. Overall, the largest mean differences tend to be at levels above (pressures lower than) ∼20 hPa prior to 1998, with mean differences as large as ±1.5% of a hemisphere; at higher pressures in the lower stratosphere where the bulk of polar processing takes place, average differences are often well within ±1% of a hemisphere during this time, and within ±0.5% of a hemisphere thereafter. Despite the better agreement, in later years, many of the differences remain statistically significant after 1998; given the low standard deviations, these results indicate small but persistent (i.e., roughly constant) differences relative to the REM.

Differences in NH A_{NAT} ~~differences in the NH~~ from the REM (Figure 8), ~~like the corresponding minimum temperature differences,~~ show more complex patterns than those in the SH and less of an obvious convergence toward better agreement after ~~1999.~~ 1998, similar to the corresponding T_{min} differences. The differences do decrease after about ~~1999 to~~ 2000, with

~~magnitudes of most differences below about 0.3~~most average differences being between ±0.25% of a hemisphere~~after 2000. JRA-55 shows a~~ . JRA-55 does show a narrow band of slightly larger ~~differences in~~ positive differences continuing into the later years between about 30 and 15 hPa. ~~CFSR and ERA-I (and to a lesser degree, MERRA) show an increase in average differences~~ MERRA-2 and ERA-I exhibit a pattern of opposing differences in this same layer between 1986 and 1998, but a layered structure of positive and negative differences at the lower levels is mostly only apparent in MERRA-2, consistent with the structure of the T_{min} differences seen in Figure 5. Overall, the differences are mostly negative in CFSR/CFSv2 and ERA-I, and positive in JRA-55 and ~~standard deviations between about 1994 and 1998.~~ MERRA-2, but there is a considerable dependence on time and pressure for all the reanalyses. As was the case for the SH, the standard deviations decrease over time ~~. CFSR and ERA-I have larger standard deviations (indicating greater variation in the differences during each year) than JRA-55 and MERRA in the first approximately half of the comparison period , but show significant decreases after about~~ with the largest values seen before 2001. There is a considerable year-to-year variability in the standard deviations at the higher levels in the earlier period with some years especially standing out (1986 in CFSR/CFSv2 and MERRA-2, 1996 in ERA-I and MERRA-2, and 2000 ~~to 2001 such that they become comparable to those for the other reanalyses. These~~ in ERA-I). These highest levels tend to be where A_{NAT} is climatologically marginal (see Figure 6).

Overall, the patterns of differences ~~are consistent with those seen~~ in A_{NAT} qualitatively follow those in T_{min} ~~. While the average differences are small throughout the comparison period, they are often significantly differentfrom zero, and there is not~~ in both hemispheres: positive/negative differences in A_{NAT} correspond to negative/positive differences in the minimum temperatures, as expected. However, the patterns of statistical significance are often different. For example, broad patches of largely statistically insignificant differences in T_{min} in ~~general a clear trend towards less significance. The patterns of significant average differences are distinctly different than those for minimum temperatures~~MERRA-2 and JRA-55 in both hemispheres after 1998 do not always translate into differences in A_{NAT} marked as not significant. Furthermore, the largest (most positive) values of one diagnostic do not always yield the smallest (most negative) ones in the other and vice versa. Even more strikingly, the patterns of standard deviations, while overall similar, do not exhibit a simple monotonic relationship with those in T_{min} and generally display much more year-to-year variability before 2000. This is not unexpected as A_{NAT} differences depend not only on overall temperature biases but also on the morphology of the fields (e.g., spatial patterns or gradients), which varies from year to year and, to a certain extent, among the reanalyses. While closely related, the A_{NAT} and T_{min} statistics represent different diagnostics and elucidate different aspects of the reanalyses' differences in relation to polar processing.

3.2 Vortex Diagnostics

Figure 9 shows the NH and SH climatologies of REM MPVG. The evolution of MPVG is quite similar in both hemispheres, particularly above 500 K; the gradients in sPV gradually increase over time, reaching maxima in roughly ~~mid-Feb~~ mid-February in the NH and early ~~Oct~~ October in the SH. These patterns largely reflect two ~~competing~~ effects: one is the seasonal cycle of the vortex building up strength and subsiding. The other is the build-up effect from wave breaking and mixing/erosion of PV in the surf zone (the region of low-magnitude PV outside the vortex, e.g., McIntyre and Palmer, 1984) over the season, which can act to sharpen the gradients of PV in the vortex edge region. ~~In the absence of large disturbances, large MPVG indicates~~ Generally, MPVG provides a measure of the strength of the vortex edge as a ~~barrier to transport~~ transport barrier. For simplicity, in the discussion of results below, we will refer to 1.0×10^{-6} s^{-1} deg^{-1} as 1 scaled PV gradient unit, or 1 PVGU.

The averages and standard deviations of differences from the REM SH MPVG are shown in Figure 10. ~~All the reanalyses~~ Through about 1998–2000, ERA-I and JRA-55 show similar patterns of differences from ~~; particularly in , and~~ the REM, with a band of near zero (for JRA-55~~before roughly 1999, there is one band of positive differences on the order of 2 – 4 PVGUs between roughly 490 and~~) or small negative (for ERA-I, magnitudes up to ∼1.5 PVGU) differences that are usually not statistically significant below about 460 K, positive differences between about 460 and 660 K, and negative differences above. The ERA-I differences are generally not statistically significant between 580 and 750 K~~that is sandwiched between negative differences at levels~~ . In the same time period, MERRA-2 shows an approximately opposite pattern, with negative differences from about 460 to 660 K and positive differences above and below~~. This positive band seems to be slightly higher in , ranging from roughly 550 to 660~~; from about 1995 through 1999, MERRA-2 shows large average differences up to about 3.5 PVGUabove 700 K. In contrast to the banded structures in the other reanalyses, CFSR/CFSv2 generally shows small magnitude negative differences across the levels and period, except during the period from 1985 through 1996 above about 700 K. The ~~standard deviations in these positive bands~~ seasonal standard deviations of the differences are relatively small (usually on the order of ~~1 – 2~~ 0.5–1.5 PVGUs), ~~indicating that the reanalyses tend to have systematically larger~~ suggesting that statistically significant differences in the reanalyses typically represent differences that are more systematic in nature for MPVG at these levels and times. The standard deviations ~~also show that the variances of the differences~~ tend to increase with height, especially at levels in the middle stratosphere above ~~700~~about 660 K where the differences ~~often~~ can exceed 2.5 – 3 PVGUs. ~~There~~ CFSR/CFSv2 and JRA-55 generally show slightly lower standard deviations than ERA-I and MERRA-2, and MERRA-2 shows a cluster of years between about 1994 and 1998 with large standard deviations above 700 K. After about 1998, there is a noticeable shift toward better agreement ~~after 1999~~ in most regions, similar to that seen in the SH temperature diagnostics~~; many of the average differences are near zero, and relatively few of the average differences at different levels and years after 1999 are significantly different from zero. This~~ . CFSR/CFSv2 and JRA-55 do not show an obvious improvement below about 550 K, but already had close agreement with the REM there. In ERA-I and MERRA-2, most regions show small (magnitude less than 1 PVGU) differences that are not statistically significant after 1998. This shift toward better agreement is also reflected in the standard deviations ~~,~~which markedly decrease in all the reanalyses, especially at levels ~~in the lower stratosphere below 660~~above about 580 K. ~~As for~~ Similar to the temperature diagnostics, the TOVS to ATOVS transition most likely played a large role in this shift, with differences in the handling of this transition and the addition of AIRS radiances in 2002 also expected to be significant factors.

~~In contrast to the temperature diagnostics shown in Section 3.1, the magnitudes of MERRA — MERRA-2 MPVG differences are generally as large as or larger than those for the other reanalyses. Two likely reasons for this are the improvement in MERRA-2 polar winter PV fields from using the cubed sphere grid (Gelaro et al., 2017) and the fact that MERRA provided PV only on a reduced resolution grid and on pressure levels with much coarser spacing than the model levels.~~

~~For MPVG in the NH~~ Differences in NH MPVG from the REM (Figure 11) ~~, there are no predominant patterns among the reanalyses.~~ indicate that CSFR/CFSv2 generally has smaller, and JRA-55 ~~both tend to have largerMPVG than~~ larger, PV gradients than the REM at levels up through about 750 K. ERA-I and MERRA-2 ~~by 1 – 3+ PVGUs across most of the levels~~ show smaller and ~~years, while tends to be smaller at most levels and years by 1 – 2 PVGUs. The~~ less systematic patterns of differences that typically are not statistically significant. ERA-I ~~average differences look relatively similar to those for the SH with a positive band~~ does show a small vertical region with significant positive differences from the REM between 520 and 580 K ~~, but overall the differences are much smaller than those in the SH, and many more are not significantly different from zero~~until about 2001, similar to its pattern for the SH but with overall smaller differences. The standard deviations of the differences are largely consistent ~~between the reanalyses, with values that~~ among the reanalyses; other than a few standout cases in ERA-I and MERRA-2 (1994/1995 in MERRA-2, and 2000/2001 in both) the standard deviations tend to increase consistently with height from less than 0.8 PVGUs at the lowest levels, to ~~above~~ about 1.5+ PVGUs ~~. All of the reanalyses show signatures of several years between 1994/95 and 2000/01 having larger magnitude average differences and standard deviations at levels above 620K, as was the case for the temperature diagnostics shown in Section 3.1.~~ at the highest levels. There is some indication of ~~a~~ convergence toward better agreement ~~after roughly 2002,~~ in MPVG after roughly 2001 in MERRA-2 and ERA-I (when the reanalyses~~(except~~ , excepting JRA-55~~)~~ , began assimilating AIRS radiances; ~~(~~Figure 1)~~. This is particularly clear in , for which most of the average differences are not significantly different from zero over all of the levels. This pattern is also apparent in and , but not for ; the differences~~ , though most differences from ~~do not seem to noticeably change after 2002.~~ the REM for these two reanalyses were not statistically significant even in the earlier years. No qualitative improvement in agreement with the REM is apparent in the CFSR/CFSv2 or JRA-55 differences, but the standard deviations of the differences do seem to decrease slightly above about 580 K for years after 2001.

Figure 12 shows the REM climatologies of SVA for both hemispheres. ~~Similar to~~ As was the case for MPVG, the seasonal patterns of SVA for both hemispheres are ~~quite~~ similar. In this case, the patterns are largely due to the lack of sunlight early in the winter season, which gradually returns later on. However, there are notable differences between the hemispheres, particularly that SVA tends to be smaller in the NH; this is because the NH polar vortex is almost always smaller than its SH counterpart. The NH also shows relatively larger values in early winter above about 650 K, resulting from the NH vortex being more often disturbed and shifted to lower latitudes within sunlight. During individual winters, and given sufficiently low temperatures, the amount of vortex air exposed to sunlight at any time is generally indicative of the amount of air where ozone depletion can take place.

The averages and standard deviations of differences ~~from SH SVA~~ of SH SVA from the REM are shown in Figure 13. ~~Here, the average differences are relatively consistent between the reanalyses, with the four reanalyses generally having bands of positive average differences(indicating the reanalyses having higher SVA) on the order of 1 – 1.5of a hemisphere in between about 520 to~~ There are some persistent patterns of differences among the reanalyses; JRA-55 SVA is consistently smaller than that from the REM between about 430 and 700 K and larger above and below. Above about 700 K ~~. These bands~~ JRA-55 differences are generally not statistically significant through about 2003, after which each of the other reanalyses evaluated had started assimilating AIRS radiances (see Figure 1 and Fujiwara et al., 2017) . The other reanalyses generally show sandwiched structures of negative and positive differences: MERRA-2 (ERA-I) shows positive (negative) values between 430 and 520 K, with negative (positive) values above and below. CFSR/CFSv2 shows positive values between about 490 and 660 K, and small (often not statistically significant) negative values at higher and lower levels; in this case, the band of positive differences ~~are consistently largest and widest for and~~ extends to higher levels after 1998. In the top several levels (approximately 750 to 850 K), agreement of the reanalyses with the REM appears to degrade starting about 1999-2000: ERA-I ~~,~~ and MERRA-2 show a decrease in the number of values that are not significantly different from zero, while JRA-55 shows a similar decrease starting around 2003–2004. MERRA-2 differences increase in magnitude in this region and time period, and those in JRA-55 change from negative to positive, while ERA-I shows increased differences, near / over 2.5%, at the highest levels in 1999–2001. CFSR/CFSv2 shows an increase in the significance of the differences at these levels after 2010 (the time of the CFSR to CFSv2 transition).

The standard deviations of the differences are the highest at levels above 660 K where they are ~~usually~~ often above 1% of a hemisphere~~(and often above 1.5 –~~. These are more pronounced in ERA-I, which shows standard deviations often ranging above 1%, with some years reaching over 2% ~~of a hemisphere) ; they are also relatively large~~ above 660 K. Some slightly larger (0.4 to 0.8%) standard deviations are also seen at the lowest levels ~~around~~ (390 and 410 K~~, which is~~), which are around the top of the subvortex region for the SH~~. A shift towards better agreement is seen most clearly in , , and after 2000; this shift is particularly evident in the standard deviation of differences , which become~~ . After 2001, the standard deviations of differences are generally less than 0.4% of a hemisphere at most of the levels between 390 and ~~850~~750 K ~~. is the only reanalysis that does not show as marked of a shift toward better agreement with – particularly at 660 and 700K, SVA is systematically higher than SVA from . Investigation~~ in all the reanalyses, suggesting a small shift towards more consistent SVA differences compared to the REM among the reanalyses. Examination of the reanalyses' differences in vortex area from those in the REM reveal they are nearly identical to ~~the ones for SVA. This indicates~~ those for SVA, indicating that the differences are largely dominated by differences in the ~~size of~~ area enclosed within the vortex edge contours.

~~Figure 14 shows the same~~ The patterns of averages and standard deviations of differences in SVA ~~, but~~ for the NH ~~. In this case, each of the reanalyses show different patterns of differences from : has a band of positive differences greater than 1of a hemisphere at the lowest levels from roughly 410 to 490K;~~ (Figure 14) are quite different than those in the SH: MERRA-2 and ERA-I ~~generally has small differences less than 1of a hemisphere at all levels except at levels around 800 to 850~~show overall positive differences (except for narrow bands of small differences at the highest levels that are not significant), while JRA-55 shows overall negative values. CFSR/CFSv2 shows negative values below about 520 K ~~; also has generally small negative differences above -1of a hemisphere; and has a band of positive differences greater than 1 – 1.5of a hemisphere betweenroughly 660 and~~ and at 800 and 850 K~~. These patterns of average differences do not noticeably change much over the full range of years, suggesting they are fairly insensitive to jumps in the observing system.~~ , with positive values in between. There is no obvious indication of a decrease in the magnitude of the differences over the period compared. The standard deviations of ~~differences from SVA all~~ NH SVA differences from the REM generally look consistent between the reanalyses, with the largest values greater than ~~1 – 1.5~~0.7–1.2% of a hemisphere usually confined to a band of levels between ~~660 to~~ 700 and 850 K. At lower levels, however, the standard deviations are quite small throughout the period, generally on the order of 0.5% of a hemisphere or less~~. Similar to SHSVA, an investigation of the reanalysis differences in vortex area from also reveals that the differences are nearly identical to the ones shown in Figure 14.~~ ; CFSR/CFSv2 shows slightly higher values below about 460 K. As was the case for the SH, the SVA differences are dominated by differences in total vortex area among the reanalyses. Thus, while there is no consistent change in agreement over the years, ~~these~~ our results indicate persistent differences in the size of the contours used to define the vortex edges, and hence some persistent differences in the isentropic PV fields (reflected in differences in the PV values at which the maximum PV gradients are located).

3.3 Derived Vortex-Temperature Diagnostics

The diagnostics shown in the following subsection are derived from the temperature and/or vortex diagnostics shown in the previous two subsections.

~~The number of days with temperatures below T_{PSC} summed over lower stratospheric levels has been previously used as a summary measure of the extent and duration of the period conducive to polar processing (e.g., Manney et al., 2011, 2015; Lawrence et al., 2~~ ~~Figure ?? shows the total days with T <~~

Figure 15 shows the winter mean volume of temperatures below T_{ice} in the SH ~~during each winter~~expressed as a fraction of the vortex volume, calculated for the “central” PSC threshold and the ±1 K sensitivity thresholds (see Section 2.2.2)~~, over pressure levels from 121.1 to 31.6hPa. The total number of SH days with T < T_{ice}, calculated from the central values, ranges from about 700 (in 1981 in MERRA and CFSR, and in 2002 in JRA-55) to over 1050 (in 1999, 2006, and 2015, and in some reanalyses in 1987 and 1996), thus showing significant interannual variability in temperatures summed over the lower stratosphere. (Note that all of the years with largest values except 1996 are years with very deep ozone holes, e.g., WMO, 2014; Nash et a~~ ~~Figures ??b and d show that these values could vary by between nearly 100 and nearly 200 days depending on the exact temperature used for the PSC threshold. The central values vary among the reanalyses by anywhere from nearly 80 days (e.g., 1980) to less than 20 days (e.g., 2006, 2011). Most of the years with the largest differences among the reanalyses are in the 1980s, with 2002, 2004, and 2012 being the only years with a spread among the reanalyses greater than 30 days after 1997. Thus there does seem to be some convergence – though not monotonic – towards better agreement among the reanalyses. There does not appear to be a consistent order of the reanalyses – the only significant pattern seems to be that ERA-I is near the low side in the period since 1998. The MERRA and MERRA-2 bars often are not next to each other, suggesting that agreement between the two is not consistently better than that among the other reanalyses. This suggests that the details of the agreement among the reanalyses depend strongly on the details of the meteorological conditions in a given year.~~

~~Figure ?? show the total number of days in the NH lower stratosphere with T < T_{NAT} each winter. The interannual variability is, of course, much larger here than in the SH, and the number of days with T < T_{NAT} often much smaller than that with T < T_{ice} in the SH. Central values range from about 120 to 250 in many of the years with strong/prolonged SSWs in December or January (1984/1985, 1998/1999, 2001/2002, 2003/2004, 2008/2009, 2012/2013) (e.g., Manney et al., 1999, 2005b, 2009, 2015; Naujokat e~~ ~~The range of values that might be seen based on the uncertainty in the PSC threshold temperature (Figure ??b and d) varies from slightly more than 100 days (e.g., 1987/1988, 2005/2006, 2012/2013) to over 300 days (e.g., 1993/1994, 2014/2015), and does not correlate strongly with the total number of days. This can be a up to about 50of the total number of days in some winters. The central number of days in an individual years varies by anywhere from 11 days (e.g., 2008) to around 80 days (e.g., 1993/1994, 1994/1995, 1995/1996), with a cluster of years from 1992/1993 to 1997/1998 with large spreads among the~~

~~reanalyses; the other large spreads between reanalyses are in 2013/2014 and 2014/2015, with most other years showing spreads between about 20 and 40 days; thus, there does not appear to be any pronounced convergence toward better agreement over the years. After about 1989, CFSR is usually near the top of the range (appearing at or near the right side of the year columns); the other reanalyses don't show obvious preferred positions. In 2011 and 2016 (the latter for the three reanalyses, ERA-I, JRA-55, and MERRA-2, that are available as we write this), the overall coldest Arctic winters in the record since 1979, the reanalyses agree very well, but not all cold winters show such close agreement (e.g., 1995 and 1996). While there are some fairly large differences and sensitivities among the reanalyses, all of the reanalyses do show similar interannual variations among the NH winters.~~

~~The winter mean volume of lower stratospheric air with temperatures below T_{PSC} (V_{PSC}) is a widely used diagnostic of polar processing potential, and it is often expressed as a fraction of the vortex volume (e.g., Rex et al., 2004, 2006; Tilmes et al., 2006; Manney et~~ ~~Here, we calculate V_{PSC} and the volume of the vortex using A_{PSC} and A_{Vort} on isentropic levels between 390 and 550K by assuming each isentropic level is nominally representative of the volume of air midway between each level; for example, the 410K level comes after 390K and before 430K, so 410K is assumed to be representative for altitudes between 400 and 420K. The altitudes for these nominal levels are determined using the Knox (1998) approximation; for the levels from 390 to 550K, this gives a mean altitude differential of 1.13km with a range of 0.98 – 1.30km. These altitude differentials are then multiplied by the area diagnostics on each isentropic level (which are converted to km^2), and summed over the vertical range to get volumes. The volume fraction is then V_{PSC}/V_{Vort}.~~

~~Figure 15 shows the winter mean volume of temperatures below T_{ice} in the SH expressed as a fraction of the vortex volume, in the same format as in Figure ??~~. Keeping in mind that T_{ice} was estimated assuming nominal pressure levels for isentropic levels (see Section 2.2.2), which ~~results in a significant overestimate~~ can result in significant overestimates of areas/volumes, Figure 15 shows that the volume fraction of cold air is relatively constant from year to year. Generally, the fractions of the vortex are between ~~0.15 and 0.25~~ 0.20 and 0.30 each year, with sensitivities to the ice threshold offsets ~~between roughly~~ often less than ±0.05~~and 0.075~~. Between the winters from 1979 and 1986, there is a very persistent pattern with CFSR/CFSv2 having the lowest, and ERA-I having the highest, cold volume fractions of the vortex. During this period, ~~can~~ CFSR/CFSv2 vortex fractions can individually be lower than the other reanalyses by nearly 0.025 to 0.03. ~~After these years , for~~ These same years also have the largest inter-reanalysis spreads, with differences between the largest and smallest vortex fractions often greater than 0.04. For nearly all years between ~~1995~~ 1996 and 2016, ERA-I ~~shows~~ tends to have the lowest volume fractions, ranging from roughly 0.01 – 0.02 lower than the other reanalyses. For years from ~~2009~~ 2007 to 2015, JRA-55 consistently has the highest volume fractions, but in these cases the ~~differences from the other reanalyses~~ inter-reanalysis differences are generally quite small. Differences among the reanalyses in the temperature threshold sensitivity envelopes ~~among the reanalyses do suggest some minor differences in horizontal temperature gradients , but nothing overtly persistent.~~ are quite small, which indicates that there are not any persistent differences in temperature gradients among the reanalyses.

Potential polar processing volumes in the NH are much lower and much more variable than those in the SH. The NH fraction of vortex volume below T_{NAT} (Figure 16) shows values in the colder years that are comparable to those below T_{ice} in the SH. The lowest values are seen in 1984/1985, 1998/1999, 2001/2002, and 2003/2004, all years with very early (mid-December to the beginning of January) major SSWs that profoundly affected the entire stratosphere, including strongly disrupting the lower stratospheric vortex (e.g., Manney et al., 1999, 2005b; Naujokat et al., 2002); in these years the fractional ~~volume is~~ volumes are near 0.03, as opposed to nearly 0.30 in the coldest years (e.g., 1996, 2011, 2016). The range of values from the PSC threshold temperature sensitivity tests varies from about ~~0.03~~ ±0.02 in the warmest years up to ~~nearly 0.10~~ over ±0.05 in the coldest years, with differences betweeen reanalyses indicating some differences in horizontal temperature gradients (especially in, e.g., ~~2011 and 2014~~1997, 2009, and 2011). The interannual variability is well represented in all of the reanalyses. The central values usually vary more between reanalyses in colder years – e.g., 1996 and 2011 ~~stands~~ stand out as showing ~~a very wide range of about 0.06. As in the SH, through 1989 CFSR stands out as having the lowest NAT volumes. CFSR and ERA-I are among the lowest during most of the record. JRA-55 typically has the largest NAT volumes during most of the period examined, with only MERRA-2 being slightly larger in 2001/2002 through 2003~~wide ranges of about 0.045. Between 1992/~~2004 and 2005~~1993 and 2016/~~2006.~~ 2017, ERA-I tends to have the smallest vortex fractions. In contrast, JRA-55 tends to have the overall largest NAT vortex fractions, having the largest values for 32 of the 38 years, with many cases being noticeably offset from the other reanalyses. While many of the recent years show smaller ranges of central values than the early years, there is not a monotonic progression, so any trend towards better agreement is masked by the larger influence of specific interannually varying conditions that affect the PSC volumes.

~~The SH vortex breakup is of considerable concern because it results in the dispersal of ozone-depleted vortex air over mid-latitudes (e.g., Ajtić et al., 2003, 2004; Manney et al., 2005c; Pazmino et al., 2005; WMO, 2007) . While ozone depletion in the Arctic has not yet been large enough for this to be an ongoing concern, vortex evolution during the 2011 Arctic vortex breakup led to significant areas of ozone depleted air over populated regions associated with increased surface UV (e.g., Manney et al., 2011; Bernhard et al., 2012) . To examine the variability and representation in reanalyses of the vortex breakup in the lower to middle stratosphere, we examine approximate vortex decay dates, which we derive from the~~ We note that the results of intercomparisons of V_{PSC}/V_{Vort} outlined above are not very sensitive to the vortex ~~area diagnostic using the $+0.1 \times 10^{-4}$ s^{-1} offsets on isentropic levels from 460 to 850K. To accomplish this, we examine NH A~~volumes. When comparing these V_{PSC}/V_{Vort} ~~between 1 Dec and 1 Jun, and SH A~~results having V_{Vort} ~~between 1 May and 1 Mar; we have defined the decay date as the last day before which A~~determined from the reanalyses' individual vortex areas to V_{PSC}/V_{Vort} ~~is above 2of a hemisphere continuously for 30 days.We choose 2of a hemisphere as the limit because it is well below the NH DJFM and SH MJJASO A~~calculated using the REM V_{Vort} ~~climatologies at all levels, which guarantees that any time the vortex is that small, it is either significantly disturbed or in the process of decaying. The 30 day limit is chosen to help guarantee that the vortex was sufficiently coherent beforehand. Our results are not highly sensitive to changing the area threshold or using vortex area with/without the sPV offset; except in some marginal cases (discussed below), adjusting the area threshold between 2 and 4only modifies the decay dates by less than 12 days in the NH, and less than 5 days in the SH.~~

(i.e., the reanalyses' V_{PSC} divided by the REM V_{Vort}), the reanalysis magnitudes and orderings remain generally consistent. However, using the REM V_{Vort} does tend to decrease the inter-reanalysis spreads and the sensitivities to the ±1 K PSC temperature thresholds. Figures 17 and 18 show pixel plots of the REM vortex decay dates and the differences from the other reanalyses (i.e., the reanalyses minus REM). For the SH, Figure 17a shows that the vortex tends to decay fairly late in the year, and it does so earlier at the upper levels than at the lower levels; in other words, the vortex in the SH typically decays from the top down. The differences in decay dates among the reanalyses are generally less than 2 weeks~~, and in most cases are between -4 and 4 days. All the reanalyses show similar patterns , at least between 1979 and 1999, with a band of positive differences for levels from roughly 580 to 660K sandwiched between bands of negative differences at the top and bottom levels. After 1999, these positive differences seem to expand upward to higher levels; this is especially the case for , which shows some of the largest positive differences in decay dates from after 2000. These results seem generally consistent with the~~ ; over 90% of the differences are between ± 7 days. The patterns of differences ~~in SVA~~ in each of the reanalyses generally follow their vortex area differences from the REM (which, while not shown, are very similar to differences shown in Figure 13~~, with positive (negative)differences in decay dates~~). That is, wherever the reanalyses have smaller/larger vortex areas than the REM, their vortices persist for less/more time. ERA-I shows some notable exceptions to this at levels above 700 K where its decay dates precede the REM by up to 14 days in the same ~~regions there are positive (negative) differences in vortex area . There are a few cases with very large positive differences , particularly in 2002 and 2009, that show up in , , and . In the 2002 SH winter, a major SSW and vortex split led to the vortex breaking down at levels above 850K by mid-October; in , the vortex area oscillated above and below 2of a hemisphere at 850K, whereas in and it stayed above 2consistently for more than 2 extra weeks. Although there was no SH SSW in 2009, similar marginal conditions occurred at 850K late in the season, leading to large differences in the decay dates in , , and~~ region where ERA-I tended to have positive vortex area differences (see, e.g., 1980, 1981, 1987). Overall, because the differences in vortex area tend to be dominant and persistent as shown in Figure 13, so too are the decay date differences, and thus there are no easily discernible changes in agreement over time.

Figure 18 shows that the NH vortex breakup is much more variable from year to year than that in the SH. Unlike the SH vortex, the NH vortex can decay nearly simultaneously over a wide range of levels (e.g., 1984 and 1999), or it can decay earlier at some low levels, and later at higher levels (e.g., 2001 and 2009). Such variability in vortex decay is due to large variability induced by SSW disturbances to the vortex, as well as polar night jet oscillation events in which the middle and upper stratospheric vortex rapidly reforms following some major and minor disturbances ~~(e.g., Hitchcock et al., 2013; Lawrence and Manney, 2018, an~~ The reanalyses' differences from the REM are generally quite small~~, usually within -2 to 2 days. There are~~ ; over 90% of the differences are between ± 4 days. With the exception of JRA-55, the reanalyses show no predominant patterns of differences (e.g., positive or negative bands)~~, but there are many more outliers~~ . JRA-55 does seem to have a slightly more pronounced band of negative differences from about 620 to 700 K (with a band of small, but positive differences above), in the same region where the JRA-55 vortex area differences tend to be the most negative (not shown directly, but consistent with Figure 14). There are also several outlier cases with absolute differences from the REM greater than 20 days (denoted by the white x symbols). ~~Many~~ Most of these cases are marginal scenarios when ~~the vortex area hovers~~ either the REM or the reanalyses' vortex areas oscillate above and below the specified ~~2~~1% of a hemisphere threshold at some levels, causing our algorithm to pick ~~a much earlier decay date than the other reanalyses that (in comparison to) persistently stay above 2at the end of the season. Similar to the SH decay dates, some of the persistent differences among the reanalyses in NH decay dates are due to the persistent differences in vortex area discussed in Section 3.2; tends to have higher vortex area than between roughly 430 and 490~~disparate decay dates. Many of these outlier cases occur at different singular levels and years in the reanalyses, but 460 K ~~, at the same~~

~~levels has some of the largest (non-outlier) decay date differences. also shows this behavior at levels between 660~~ 2003/2004 does show up as a negative outlier in both CFSR/CFSv2 and ~~800~~JRA-55, while 660 K 2005/2006 shows up as a positive outlier in both CFSR/CFSv2 and ERA-I.

4 Conclusions

We have herein done an extensive intercomparison of diagnostics ~~of~~ relevant to polar chemical processing among ~~the five most~~ four recent full-input reanalyses, using the ~~most recent, MERRA-2,~~ "reanalysis ensemble mean" (REM) as a reference to compare ~~MERRA, ERA-I, JRA-55, and CFSR~~CFSR/CFSv2, ERA-I, JRA-55, and MERRA-2. The diagnostics we compare are based on polar vortex and temperature conditions in the ~~lower~~ lower-to-middle stratosphere, and comprise measures of PSC formation and chlorine activation based on temperatures; vortex size, strength, and sunlight exposure; and additional diagnostics derived from those directly obtained from temperatures and vortex characteristics. They thus provide a thorough assessment of the reanalyses' representation of the potential for polar processing and ozone loss in both hemispheres. ~~Compared to previous studies, we include all of the latest generation reanalyses, examine the sensitivity of the diagnostics to uncertainties in temperature and vortex threshold values used, and provide an assessment of the statistical significance of the differences between reanalyses.~~ The main findings of our analyses are summarized in the following subsection.

4.1 Summary

Temperature diagnostics related to polar processing converge towards better agreement in the SH over the period compared (from 1979 to present)~~, with agreement~~ . In the period prior to ∼1999, reanalysis differences in minimum temperatures ~~generally within about 1~~compared to the REM could be as large as ±3 K~~after 1998 (as opposed to up to about 6~~, particularly at pressures below 30 hPa; in years after, reanalysis minimum temperature differences from the REM decrease to within roughly ±0.5 K ~~before) and largest~~ throughout the 120–10 hPa column. The reanalysis differences from the REM for SH areas with temperatures below the NAT PSC threshold (A_{NAT}~~differences decreasing from near 2~~) show a similar and consistent shift, with differences among the reanalyses being as large as ±1.5% ~~to less than about~~ of a hemisphere prior to ∼1999, but within ±0.5% of ~~the hemisphere . A large sudden decrease in both the reanalysis differences and the standard deviations thereof is seen in 1999,~~ a hemisphere thereafter. This shift toward better agreement in ∼1999 is seen as both a sudden decrease among the winter-averaged differences from the REM, and as a sudden decrease in standard deviations of reanalysis minus REM differences, which is consistent with previous studies (e.g., Long et al., 2017) that show large improvements in zonal mean temperatures after the ~~TOVS/ATOVStransition~~reanalyses transition from assimilating TOVS observations to including ATOVS. In the NH, the agreement ~~before 1998~~ among the reanalyses before ∼1999 was already much closer (~~within about 2~~generally within ±1.5 K from the REM for minimum temperatures, but often within a much smaller margin at pressures greater than 30 hPa), but the average differences and standard deviations ~~do decrease to some extent over the years. Differences in both hemispheres show a banded structure with height , with generally being warmer than~~ also decreased to a lesser extent thereafter. The structure of average differences, particularly before ∼1999, is varied among the reanalyses. MERRA-2 and

ERA-I generally showed banded structures of average differences from the REM that changed signs with height prior to ∼1999. CFSR/CFSv2 tended to have average differences of the same sign throughout the 120–10 hPa column up until 1987, after which the differences switched signs (as in the case of minimum temperatures) or became more varied (as in the case of A_{NAT}). The structure of average differences from the REM for JRA-55 was generally a bit more complicated than that in the other reanalyses~~between about 50 and 30~~, but did show that the signs of the differences changed in the lower stratosphere between 100–30 hPa~~; the~~. The standard deviations of ~~the differences generally increase with height . Between about 1994 and 1999, increased differences and standard deviations are seen above about~~ differences from the REM were quite consistent among the reanalyses; they increased with height (not necessarily monotonically), particularly at pressures lower than 30 hPa~~in both hemispheres, which may be related to increasing impacts of differences in how the data are assimilated.~~.

~~In both hemispheres, temperatures diagnostics before 1998 show closer agreement between MERRA-2 and MERRA than between MERRA-2 and the other reanalyses.~~

Differences from the REM among the reanalyses for SH maximum PV gradients ~~are generally similar, with banded structures of positive and negative differences from (particularly in CFSR, , and) , and standard deviations that increase with height. The differences from in these cases are generally within~~ (MPVG) showed a similar convergence toward better agreement as did the temperature diagnostics. Differences from the REM were within roughly ±~~4~~ 2.5 PVGUs prior to ∼1999, but ~~only roughly~~ within ±1 ~~PVGU thereafter. For NH maximum PV gradients, there are no consistent patterns of differences from among the reanalyses, but, similar to the SH , the standard deviations of the differences tend to increase with height . Generally the differences are within ±~~PVGUs after. The standard deviations of the SH differences increased with height to values that were commonly above 2 ~~PVGUs over all the years; however, there is a noticeable convergence in agreement between~~ PVGUs, particularly at isentropic levels above 600 K; after 1999, these standard deviations decreased in magnitude, but the pattern of values increasing with height remained consistent. In these cases the differences from the REM for SH MPVG were consistently negative across all years for CFSR/CFSv2, while those for JRA-55 were consistently positive. ERA-I and MERRA-2 ~~after 2002 that is not as apparent in the other reanalyses. Differences from in sunlit vortex area generally follow differences in vortex area itself. In the SH these average SVA differences can be as large as 2of a hemisphere, but they decrease in magnitude (in both the averages and standard deviations) after roughly~~ had banded structures of differences similar to those in the SH temperature diagnostics that mostly disappeared after ∼1999. In the ~~NH the average differences are generally small, except for some persistent positive average differences in limited bands that show up in between roughly 660 and 750K, and in between roughly 410 and 490K. For these PV-based diagnostics, MERRA and MERRA-2 show differences of magnitude as large as (sometimes larger than) those between MERRA-2 and the other reanalyses; the reduced resolution PV provided for MERRA and the cubed-sphere grid used for MERRA-2 are likely factors in these differences~~ case of NH MPVG, differences remained largely constant over time and potential temperature levels, generally being within ±1.5 PVGUs of the REM with standard deviations that increased with height. Here, again, CFSR/CFSv2 had average differences from the REM that were consistently negative, while JRA-55 was consistently positive. Differences from the REM in sunlit vortex area across the reanalyses in both hemispheres remained relatively constant over time, and they overall followed differences in the raw vortex areas.

~~The derived winter summary diagnostics, which include the number of days (summed over vertical levels) below PSC thresholds, the~~

In the SH, all the reanalyses showed similar magnitudes and temperature threshold sensitivities in the winter mean volume of air ~~below PSC thresholds (expressed~~ (as a fraction of ~~the~~ vortex volume) ~~, and vortex decay dates, generally agree better in the SH than in the NH. Particularly for the number of days below T_{ice} and V_{ice} for the SH, all the reanalyses show similar magnitudes and sensitivities to the ±1K temperature offsets. For~~ below ice PSC thresholds. In the NH, the ~~number of days below T_{NAT} and V_{NAT} vary~~ winter mean volume of air below NAT PSC thresholds varied much more from year to year, and the differences among the reanalyses and sensitivities to the temperature offsets ~~are~~ were much larger percentages of the actual derived values. These characteristics are in many ways to be expected, since SH winters are much more consistent from year to year than NH winters; thus, even though the individual temperature polar processing diagnostics ~~show~~ showed much larger average differences and standard deviations ~~for~~ in the SH, the aggregation of the full winter seasons ~~done for the derived diagnostics leads~~ in the winter mean V_{PSC}/V_{Vort} led to more consistent results. ~~These findings are also consistent with~~ For the vortex decay dates, ~~which, except in rare cases, generally vary by less than a week~~ the reanalyses' differences from the REM generally followed their differences from the REM in vortex area (and as a result, sunlit vortex area) in that wherever the reanalyses had larger/smaller vortex areas, they also had later/earlier vortex decay dates. The agreement among the reanalyses for ~~the SH. While the differences in decay dates are also often quite small~~ vortex decay dates was generally best in the NH, ~~the early vortex breakup (relative to the SH) and the frequent occurrence of midwinter SSWs and significant vortex disturbancesmake large differences more common because of more frequent marginal cases~~despite there being some marginal cases with large differences due to vortex disturbances.

4.2 Implications

The results shown herein illustrate some implications that may be expected for polar processing studies using reanalysis temperatures and PV in the stratosphere. These implications will generally depend on the hemisphere in question and the ~~detail~~ details of the study. For example, the derived diagnostics in Section 3.3 demonstrate that in the aggregate most SH winters in the satellite era are quite similar, and that the sensitivities to different PSC temperature thresholds are consistent among the reanalyses. However, the differences shown in Section 3.1 indicate that differences can depend strongly on the levels and years examined, especially prior to 1999 before the assimilation of AMSU data in the reanalyses. Thus, studies that discuss SH winter conditions in aggregate are less likely to be affected than detailed studies (e.g., those making use of nudged and specified dynamics models, and/or Lagrangian transport models), whose conclusions could be significantly altered by the details of how, when, and where the temperatures differ among the reanalyses. In contrast, for the NH, Section 3.1 showed that temperature diagnostic differences were relatively small among the reanalyses, but the results in Section 3.3 showed that the aggregate derived diagnostics vary widely between reanalyses in some cases, and can be highly sensitive to the specific temperature thresholds used. Clearly polar processing potential is often much smaller in the NH than in the SH, and thus conclusions based on the often marginal conditions of the NH are much more likely to be affected by small differences among the reanalyses. Thus, both detailed and aggregate studies of NH polar processing could in some cases be markedly affected by differences among the reanalyses. However, all of the reanalyses do show similar interannual variations among the derived diagnostics, and thus for purposes of putting some NH winters into the context of others (e.g., comparing how cold some are relative to others), any of the reanalyses would give similar results. The extent to which different kinds of studies of NH and/or SH polar processing may be affected is beyond the scope of this paper, but work is in progress within S-RIP to explore some of these implications.

It is difficult to assess the potential implications of differences among the reanalyses in the vortex diagnostics. Since MERRA-2 is the only reanalysis that provides PV fields on its model levels, we have applied the strategy we think other data users requiring PV on model levels would use, which was to derive PV from each reanalysis using their available model level products. Thus, it is important to recognize that the vortex diagnostics used herein are derived from PV fields that are calculated from the different reanalyses in different ways, which makes it problematic to assess whether and the extent to which the reanalysis differences are due to differences in calculations, dynamics, vertical and/or horizontal resolution, etc. Because MERRA-2 includes PV calculated within its DAS, we generally consider MERRA-2 PV to be more consistent and complete than the PV fields derived from the other reanalyses' model level data. ~~Significant differences between MERRA-2 and MERRA PV-based diagnostics suggest that resolution may be a significant factor (since the MERRA PV was also calculated within the DAS, but was only available on a reduced resolution grid), but model changes may also play a role.~~ Despite these complicating factors, our treating each of the reanalyses equally (same procedure for calculating MPVG, and ~~same contours used~~ using each of the reanalyses' climatological MPVGs to define vortex ~~area~~edges) allows us to draw some useful conclusions: While there were some small indications of convergence toward better agreement in ~~maximum PV gradients and sunlit vortex area~~ MPVG for both hemispheres among some reanalyses (see Section 3.2), there were ~~persistent differences elsewhere~~primarily persistent differences in SVA. Given the combination of differences in MPVG and SVA ~~, these results indicate that using the same vortex edge values for each reanalysis is not always be an appropriate simplification. Differences among the reanalyses in MPVG alone~~ (and raw vortex area), the results shown here indicate that there ~~may be~~ are some inherent differences in the ~~equivalent latitude mapping of the PV fields (which, again, could arise for numerous reasons)~~PV fields that lead to somewhat disparate equivalent latitude mappings, which in some cases could alter conclusions drawn about transport barriers and trace gases in equivalent latitude coordinates.

It is also possible that results for the SH were contaminated by the presence of double-peaked (bifurcated) PV gradients (e.g., Conway et al., 2018) that could have different magnitudes or structures among the reanalyses.

4.3 Recommendations

All of the reanalyses used here represent vast improvements over those commonly used a decade ago, and with those improvements comes much closer agreement in the polar processing diagnostics presented here. The older reanalyses, especially ERA-40 and NCEP/NCAR and NCEP/DOE, have long been obsolete and are not recommended for studies focused on polar processing and the stratosphere in general (see Fujiwara et al., 2017, and references therein). Any of the modern reanalyses evaluated herein are much better choices for polar processing studies as they all provide more accurate and similar representations of interannual variability in polar processing diagnostics in both hemispheres.

In general, it is always better to use more than one reanalysis, even for studies involving recent winters where it can reasonably be expected that differences among the reanalyses will be small. One of the best ways to express uncertainty in results is using multiple reanalyses, and explicitly showing and discussing how they agree/disagree, and whether any differences affect the findings; this is especially important for diagnostics that cannot be compared with observations. As previously shown by Lawrence et al. (2015) for polar processing diagnostics, and Long et al. (2017) for zonal means, our intercomparisons (see particularly Figures ~~5~~ 4 – ~~6 and 8~~ 5 and 7 – ~~9~~8) show that there are substantial (especially large in the SH) changes in temperature-based diagnostics that are clearly related to changes in assimilated data inputs among the reanalyses. Since many of the major changes in data inputs are made at approximately the same time in each reanalysis, the agreement or lack thereof between the reanalyses does not provide the information to assess the degree to which these changes are caused by changes to the assimilated observations. We thus emphasize here that reanalysis temperatures, especially in the Antarctic, are not generally suitable for assessment of trends in temperature-based diagnostics; use of reanalyses in trend studies should be regarded with skepticism and only attempted ~~with the use of multiple reanalyses, and~~ after rigorous assessment of the relationships of temperature changes to observations assimilated (which, to our knowledge, has not been done ~~).~~ for most of the reanalyses considered here); were such a study to be done, agreement among multiple reanalyses would in addition be required to consider any trends robust.

When using multiple reanalyses, it is important to treat them as fairly and equally as possible to reduce the uncertainty in sources of differences. For example, using one ~~reanalyses~~ reanalysis with data on model levels, and another ~~reanalysis~~ one with data on pressure levels is not recommended. It is also important to be clear whether and how fields/quantities are derived from the products provided by the reanalyses, as we have done herein with PV. Until and unless reanalysis centers provide standard sets of products on standardized isobaric and isentropic levels, users of reanalysis data will generally be best served by using model data to vertically interpolate and derive fields as needed. Numerous evaluations of reanalyses for S-RIP are finding, as we have here, that it would be valuable to have PV on model levels available in future reanalyses.

With regard to more specific polar processing applications: we also recommend that trends, correlations, and/or other similar analyses of diagnostics that assess low temperatures aggregated over winter months, seasons, and/or vertical levels in the NH polar region be performed with caution. ~~Figures 17 and 19 demonstrate~~ Figure 16 demonstrates that there is non-negligible interannual variability in the sensitivity to the specific temperature values chosen to represent NAT PSC thresholds that are used to calculate the ~~number of days and~~ volume of air below NAT thresholds in the NH, especially relative to the SH (~~Figures 16 and 18~~Figure 15) in which we used the lower ice PSC thresholds. The vortex diagnostics in Section 3.2 show some differences that appear to be related to biases between PV in the reanalyses, arguing for careful assessment of the sensitivity of vortex diagnostics to exact PV values. Because many of the diagnostics that are most informative about lower stratospheric polar chemical processing cannot be readily validated by comparison with data, the comparison of reanalyses is a powerful tool for assessing robustness and uncertainty in these diagnostics.

Appendix A: The HNO_3 and H_2O profiles used for the calculation of NAT and ice PSC thresholds are given in Table A1. These values were derived from Upper Atmosphere Research Satellite (UARS) measurements from the Cryogenic Limb Array Etalon Spectrometer for HNO_3 and Microwave Limb Sounder for H_2O, by averaging values for December/January 1991/1992 and 1992/1993, as described by Manney et al. (2003, 2005a) . The values are thus close to climatological values for the NH and for the SH during early winter. The values are defined at six levels per decade in pressure (standard UARS levels), and interpolated to the same 12 level per decade pressures that the reanalyses are interpolated to. Approximately corresponding isentropic surfaces were estimated by averaging climatological (Fleming et al., 1990) temperature profiles for January, April, July, and October (the solstice and equinox months) at ±70, ±50, and ±10° latitude; the potential temperatures thus derived are then adjusted to "nice" values. The pressure levels used for the analysis, their approximately corresponding potential temperature levels, and the NAT and ice PSC thresholds calculated from them are listed in Table A2. These HNO_3 and H_2O profiles, and the PSC thresholds derived from them, may not be the most appropriate for studies requiring precise estimates of the PSC thresholds and PSC potential throughout the winter. Especially in the SH, denitrification and dehydration can significantly change the profiles of HNO_3 and H_2O away from the climatology. Furthermore, using the same PSC thresholds from pressure surfaces on potential temperature surfaces is an additional approximation that tends to overestimate the size of regions with low temperatures. Regardless, these values are appropriate for defining regions and periods of time when polar processing can take place, and for understanding the differences among reanalyses.

Data availability. The datasets used/produced are publicly available, as follows:

- MERRA-2: https://disc.sci.gsfc.nasa.gov/uui/datasets?keywords=%22MERRA-2%22
- ERA-I: http://apps.ecmwf.int/datasets/
- JRA-55: Through NCAR RDA at http://dx.doi.org/10.5065/D6HH6H41
- CFSR/CFSv2 model level data: Available upon request from Karen H Rosenlof (karen.h.rosenlof@noaa.gov)
- Polar processing diagnostic products: Contact Zachary D Lawrence (zachary.lawrence@student.nmt.edu)

Author contributions. ZDL and GLM designed the study. ZDL did the analysis. KW provided expertise and guidance on the reanalysis datasets. ZDL and GLM wrote the discussion paper and KW commented on and edited it; ZDL, KW, and GLM revised the paper.

Competing interests. The authors declare no competing interests.

Acknowledgements. We thank the Microwave Limb Sounder team at JPL, especially Brian W Knosp and Ryan A Fuller, for computational, data processing, management, and analysis support; Will McCarty and Larry Coy of the GMAO for helpful comments; NASA's GMAO, ECMWF, JMA, and NCEP for providing their assimilated data products; ~~and~~ Amy Butler, Jeremiah Sjoberg, Craig Long, Sean Davis, Henry L Miller, and Karen Rosenlof for processing and providing the model level CFSR~~data~~/CFSv2 data; and Simon Chabrillat and an anonymous reviewer for their constructive and helpful comments that improved this paper. GLM and ZDL were supported by the JPL Microwave Limb Sounder team under JPL subcontracts to NWRA and NMT; KW was supported by NASA's Modeling, Analysis and Prediction (MAP) program, which also provides support for MERRA and ~~MERRA-2~~MERRA-2.

References

Ajtić, J., Connor, B. J., Randall, C. E., Lawrence, B. N., Bodeker, G. E., Rosenfield, J. E., and Heuff, D. N.: Antarctic air over New Zealand following vortex breakdown in 1998, Annales Geophysicae, 21, 2175–2183, 2003.

Ajtić, J., Connor, B. J., Lawrence, B. N., Bodeker, G. E., Hoppel, K. W., Rosenfield, J. E., and Heuff, D. N.: Dilution of the Antarctic ozone hole into southern midlatitudes, 1998-2000, J. Geophys. Res., 109, D17107, https://doi.org/10.1029/2003JD004500, 2004.

Albers, J. R. and Nathan, T. R.: Ozone Loss and Recovery and the Preconditioning of Upward-Propagating Planetary Wave Activity, J. Atmos. Sci., 70, 3977–3994, https://doi.org/10.1175/JAS-D-12-0259.1, https://doi.org/10.1175/JAS-D-12-0259.1, 2013.

Andrews, D. G.: Some comparisons between the middle atmosphere dynamics for the southern and northern hemispheres, Pure and Appl. Geophys., 130, 213–232, 1989.

Bernhard, G., Manney, G., Fioletov, V., Grooß, J.-U., Heikkila, A., Johnson, B., Koslela, T., Lakkala, K., Müller, R., Myhre, C., and Rex, M.: [The Arctic] Ozone and UV Radiation, [in “State of the Climate in 2011”.], Bull. Am. Meteor. Soc., 93, S129–S132, 2012.

Bloom, S. C., Takacs, L. L., da Silva, A. M., and Ledvina, D.: Data assimilation using incremental analysis updates, Mon. Weather Rev., 124, 1256–1271, 1996.

Boccara, G., Hertzog, A., Basdevant, C., and Vial, F.: Accuracy of NCEP/NCAR reanalyses and ECMWF analyses in the lower stratosphere over Antarctica in 2005, J. Geophys. Res., 113, n/a–n/a, https://doi.org/10.1029/2008JD010116, http://dx.doi.org/10.1029/2008JD010116, d20115, 2008.

Bosilovich, M., Akella, S., Coy, L., Cullather, R., Draper, C., Gelaro, R., Kovach, R., Liu, Q., Molod, A., Norris, P., Wargan, K., Chao, W., Reichle, R., Takacs, L., Vikhliaev, Y., Bloom, S., Collow, A., Firth, S., Labow, G., Partyka, G., Pawson, S., Reale, O., Schubert, S. D., and Suarez, M.: MERRA-2: Initial Evaluation of the Climate, Series on Global Modeling and Data Assimilation, NASA/TM–2015-104606, Vol. 43, NASA, 2015.

Bosilovich, M. G., Lucchesi, R., and Suarez, M.: MERRA-2: File Specification, Office Note 9, GMAO Office Note, 73 pp, available from http://gmao.gsfc.nasa.gov/pubs/office_notes., 2016.

Butchart, N. and Remsberg, E. E.: The area of the stratospheric polar vortex as a diagnostic for tracer transport on an isentropic surface, J. Atmos. Sci., 43, 1319–1339, 1986.

Chen, Y., Weng, F., Han, Y., and Liu, Q.: Validation of the Community Radiative Transfer Model by using CloudSat data, J. Geophys. Res., 113, n/a–n/a, https://doi.org/10.1029/2007JD009561, http://dx.doi.org/10.1029/2007JD009561, d00A03, 2008.

Conway, J., Bodeker, G., and Cameron, C.: Bifurcation of potential vorticity gradients across the Southern Hemisphere stratospheric polar vortex, Atmos. Chem. Phys., 18, 8065–8077, https://doi.org/10.5194/acp-18-8065-2018, https://www.atmos-chem-phys.net/18/8065/2018/, 2018.

Davies, S. et al.: Modeling the effect of denitrification on Arctic ozone depletion during winter 1999/2000, J. Geophys. Res., 108, 8322, https://doi.org/10.1029/2001JD000445, 2003.

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. Meteorol. Soc., 137, 553–597, 2011.

DiCiccio, T. J. and Efron, B.: Bootstrap confidence intervals, Statist. Sci., 11, 189–228, https://doi.org/10.1214/ss/1032280214, https://doi.org/10.1214/ss/1032280214, 1996.

Dunkerton, T. J. and Delisi, D. P.: Evolution of potential vorticity in the winter stratosphere of January-February 1979, J. Geophys. Res., 91, 1199–1208, 1986.

Ebita, A. et al.: The Japanese 55-year Reanalysis "JRA-55": An interim report, SOLA, 7, 149–152, 2011.

Feng, W., Chipperfield, M. P., Roscoe, H. K., Remedios, J. J., Waterfall, A. M., Stiller, G. P., Glatthor, N., Höpfner, M., and Wang, D.-Y.: Three-dimensional model study of the Antarctic ozone hole in 2002 and comparison with 2000, J. Atmos. Sci., 62, 822–837, 2005.

Fleming, E., Chandra, S., Barnett, J. J., and Corney, M.: Zonal mean temperature, pressure, zonal wind and geopotential height as functions of latitude, Adv. Sp. Res., 10, 11–53, 1990.

Forster, P. M. and Shine, K. P.: Radiative forcing and temperature trends from stratospheric ozone changes, J. Geophys. Res., 102, 10,841–10,855, 1997.

Fujiwara, M., Wright, J. S., Manney, G. L., Gray, L. J., Anstey, J., Birner, T., Davis, S., Gerber, E. P., Harvey, V. L. a nd Hegglin, M. I., Homeyer, C. R., Knox, J. A., Krüger, K., Lambert, A., Long, C. S., Martineau, P., Monge-Sanz, B. M., Santee, M. L., Tegtmeier, S., Chabrillat, S., Tan, D. G. H., Jackson, D. R., Polavarapu, S., Compo, G. P., Dragani, R., Ebisuzaki, W., Haraďa, Y., Kobayashi, C., McCarty, W., Onogi, K., Pawson, S., Simmons, A., Wargan, K., Whitaker, J. S., and Zou, C.-Z.: Introduction to the SPARC Reanalysis Intercomparison Project (S-RIP) and overview of the reanalysis systems, Atmos. Chem. Phys., 17, 1417–1452, https://doi.org/10.5194/acp-17-1417-2017, www.atmos-chem-phys.net/17/1417/2017/, 2017.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., nov, A. D., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., n, J. E. N., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version-2 (MERRA-2), J. Clim., 30, 5419–5454, https://doi.org/doi:10.1175/JCLI-D-16-0758.1, 2017.

Global Modeling and Assimilation Office (GMAO): MERRA-2 inst3_3d_asm_Nv: 3d, 3-Hourly,Instantaneous, Model-Level, Assimilation, Assimilated Meteorological Fields V5.12.4, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed 1 November 2015, https://doi.org/10.5067/WWQSXQ8IVFW8, 2015.

Global Modeling and Assimilation Office (GMAO): Use of MERRA-2 for Atmospheric Chemistry and Transport Studies,, https://gmao.gsf.nasa.gov/reanalysis/MERRA-2/docs/ANAvsASM.pdf, 2017.

Gobiet, A., Kirchengast, G., Manney, G. L., Borsche, M., Retscher, C., and Stiller, G.: Retrieval of temperature profiles from CHAMP for climate monitoring: Intercomparison with Envisat MIPAS and GOMOS and different atmospheric analyses, Atmos. Chem. Phys., 7, 3519–3536, 2007.

Han, Y., van Delst, P., Liu, Q., Weng, F., Yan, B., Treadon, R., and Derber, J.: JCSDA Community Radiative Transfer Model (CRTM)–Version 1, Tech. Rep. 122, NOAA, available online at https://docs.lib.noaa.gov/noaa_documents/NESDIS/TR_NESDIS/TR_NESDIS_122.pdf, 2006.

Hanson, D. and Mauersberger, K.: Laboratory studies of the nitric acid trihydrate: Implications for the south polar stratosphere, Geophys. Res. Lett., 15, 855–858, 1988.

Hegglin, M. I., Boone, C. D., Manney, G. L., and Walker, K. A.: A global view of the extratropical tropopause transition layer (ExTL) from Atmospheric Chemistry Experiment Fourier Transform Spectrometer O_3, H_2O, and CO, J. Geophys. Res., 114, D00B11, https://doi.org/10.1029/2008JD009984, 2009.

Hitchcock, P., Shepherd, T. G., and Manney, G. L.: Statistical characterization of Arctic polar-night jet oscillation events, J. Clim., 26, 2096–2116, 2013.

Hoffmann, L. and Alexander, M. J.: Retrieval of stratospheric temperatures from Atmospheric Infrared Sounder radiance measurements for gravity wave studies, J. Geophys. Res., 114, https://doi.org/10.1029/2008JD011241, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JD011241, 2009.

Hoffmann, L., Hertzog, A., Rößler, T., Stein, O., and Wu, X.: Intercomparison of meteorological analyses and trajectories in the Antarctic lower stratosphere with Concordiasi superpressure balloon observations, Atmos. Chem. Phys., 17, 8045–8061, https://doi.org/10.5194/acp-17-8045-2017, https://www.atmos-chem-phys.net/17/8045/2017/, 2017.

Huck, P. E., McDonald, A. J., Bodeker, G. E., and Struthers, H.: Interannual variability in Antarctic ozone depletion controlled by planetary waves and polar temperatures, Geophys. Res. Lett., 32, L13819, https://doi.org/10.1029/2005GL022943, 2005.

Knox, J. A.: On converting potential temperature to altitude in the middle atmosphere, Eos Trans. AGU, 79, 376, 1998.

Knudsen, B. M., Rosen, J. M., Kjome, N. T., and Whitten, A. T.: Comparison of analyzed stratospheric temperatures and calculated trajectories with long-duration balloon data, J. Geophys. Res., 101, 19,137–19,145, 1996.

Knudsen, B. M., Pommereau, J.-P., Garnier, A., Nunez-Pinharanda, M., Denis, L., Letrenne, G., Durand, M., and Rosen, J. M.: Comparison of stratospheric air parcel trajectories based on different meteorological analyses, J. Geophys. Res., 106, 3415–3424, 2001.

Knudsen, B. M. et al.: Accuracy of analyzed stratospheric temperatures in the winter Arctic vortex from infrared Montgolfier long-duration balloon flights 2. Results, J. Geophys. Res., 107, D001329, 2002.

Kobayashi, S., Matricardi, M., Dee, D., and Uppala, S.: Toward a consistent reanalysis of the upper stratosphere based on radiance measurements from SSU and AMSU-A, Q. J. R. Meteorol. Soc., 135, 2086–2099, 2009.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specification and Basic Characteristics, J. Meteor. Soc. Japan, 93, https://doi.org/10.2151/jmsj.2015-001, 2015.

Lacis, A., Wuebbles, D. J., and Logan, J. A.: Radiative forcing of climate by changes in the vertical distribution of ozone, J. Geophys. Res., 95, 9971–9981, 1990.

Lahiri, S.: Resampling Methods for Dependent Data, Springer Series in Statistics, Springer, 2003.

Lambert, A. and Santee, M. L.: Accuracy and precision of polar lower stratospheric temperatures from reanalyses evaluated from A-Train CALIOP and MLS, COSMIC GPS RO, and the equilibrium thermodynamics of supercooled ternary solutions and ice clouds, Atmos. Chem. Phys., 18, 1945–1975, https://doi.org/10.5194/acp-18-1945-2018, https://www.atmos-chem-phys.net/18/1945/2018/, 2018.

Lawrence, Z. D. and Manney, G. L.: Characterizing Stratospheric Polar Vortex Variability With Computer Vision Techniques, Journal of Geophysical Research: Atmospheres, 123, 1510–1535, https://doi.org/10.1002/2017JD027556, http://dx.doi.org/10.1002/2017JD027556, 2017JD027556, 2018.

Lawrence, Z. D., Manney, G. L., Minschwaner, K., Santee, M. L., and Lambert, A.: Comparisons of polar processing diagnostics from 34 years of the ERA-Interim and MERRA reanalyses, Atmos. Chem. Phys., 15, 3873–3892, 2015.

Levine, J. G., Braesicke, P., Harris, N. R., Savage, N. H., and Pyle, J. A.: Pathways and timescales for troposphere-to-stratosphere transport via the tropical tropopause layer and their relevance for very short lived substances, J. Geophys. Res., 112, D04308, https://doi.org/10.1029/2005JD006940, 2007.

Livesey, N. J., Read, W. G., Wagner, P. A., Froidevaux, L., Lambert, A., Manney, G. L., Millán Valle, L. F., Pumphrey, H. C., Santee, M. L., Schwartz, M. J., Wang, S., Fuller, R. A., Jarnot, R. F., Knosp, B. W., and Martinez, E.: EOS MLS Version 4.2x Level 2 data quality and description document, Tech. rep., JPL, available from http://mls.jpl.nasa.gov/, 2015.

Long, C. S., Fujiwara, M., Davis, S., Mitchell, D., and Wright, C.: SPARC Reanalysis Intercomparison Project (S-RIP) Final Report. Chapter 3: Overview of Winds and Temperatures, in preparation, to be published by SPARC in 2018.

Long, C. S., Fujiwara, M., Davis, S., Mitchell, D. M., and Wright, C. J.: Climatology and Interannual Variability of Dynamic Variables in Multiple Reanalyses Evaluated by the SPARC Reanalysis Intercomparison Project (S-RIP), Atmos. Chem. Phys., 17, 14,593–14,629, https://doi.org/10.5194/acp-17-14593-2017, 2017.

Manney, G. L. and Lawrence, Z. D.: The major stratospheric final warming in 2016: Dispersal of vortex air and termination of Arctic chemical ozone loss, Atmos. Chem. Phys. Disc., 16, https://doi.org/10.5194/acp-2016-633, http://www.atmos-chem-phys-discuss.net/acp-2016-633/, 2016.

Manney, G. L., Zurek, R. W., Gelman, M. E., Miller, A. J., and Nagatani, R.: The anomalous Arctic lower stratospheric polar vortex of 1992–1993, Geophys. Res. Lett., 21, 2405–2408, 1994.

Manney, G. L., Swinbank, R., Massie, S. T., Gelman, M. E., Miller, A. J., Nagatani, R., O'Neill, A., and Zurek, R. W.: Comparison of U. K. Meteorological Office and U. S. National Meteorological Center stratospheric analyses during northern and southern winter, J. Geophys. Res., 101, 10,311–10,334, 1996.

Manney, G. L., Lahoz, W. A., Swinbank, R., O'Neill, A., Connew, P. M., and Zurek, R. W.: Simulation of the December 1998 stratospheric major warming, Geophys. Res. Lett., 26, 2733–2736, 1999.

Manney, G. L., Sabutis, J. L., Pawson, S., Santee, M. L., Naujokat, B., Swinbank, R., Gelman, M. E., and Ebisuzaki, W.: Lower stratospheric temperature differences between meteorological analyses in two cold Arctic winters and their impact on polar processing studies, J. Geophys. Res., 108, 8328, https://doi.org/10.1029/2001JD001149, 2003.

Manney, G. L., Allen, D. R., Krüger, K., Naujokat, B., Santee, M. L., Sabutis, J. L., Pawson, S., Swinbank, R., Randall, C. E., Simmons, A. J., and Long, C.: Diagnostic Comparison of Meteorological Analyses during the 2002 Antarctic Winter, Mon. Weather Rev., 133, 1261–1278, 2005a.

Manney, G. L., Krüger, K., Sabutis, J. L., Sena, S. A., and Pawson, S.: The remarkable 2003-2004 winter and other recent warm winters in the Arctic stratosphere since the late 1990s, J. Geophys. Res., 110, D04107, https://doi.org/10.1029/2004JD005367, 2005b.

Manney, G. L., Santee, M. L., Livesey, N. J., Froidevaux, L., Read, W. G., Pumphrey, H. C., Waters, J. W., and Pawson, S.: EOS Microwave Limb Sounder observations of the Antarctic polar vortex breakup in 2004, Geophys. Res. Lett., 32, L12811, https://doi.org/10.1029/2005GL022823, 2005c.

Manney, G. L., Daffer, W. H., Zawodny, J. M., Bernath, P. F., Hoppel, K. W., Walker, K. A., Knosp, B. W., Boone, C., Remsberg, E. E., Santee, M. L., Harvey, V. L., Pawson, S., Jackson, D. R., Deaver, L., McElroy, C. T., McLinden, C. A., Drummond, J. R., Pumphrey, H. C., Lambert, A., Schwartz, M. J., Froidevaux, L., McLeod, S., Takacs, L. L., Suarez, M. J., Trepte, C. R., Cuddy, D. C., Livesey, N. J., Harwood, R. S., and Waters, J. W.: Solar occultation satellite data and derived meteorological products: Sampling issues and comparisons with Aura Microwave Limb Sounder, J. Geophys. Res., 112, https://doi.org/10.1029/2007JD008709, 2007.

Manney, G. L., Schwartz, M. J., Krüger, K., Santee, M. L., Pawson, S., Lee, J. N., Daffer, W. H., Fuller, R. A., and Livesey, N. J.: Aura Microwave Limb Sounder observations of dynamics and transport during the record-breaking 2009 Arctic stratospheric major warming, Geophys. Res. Lett., 36, L12815, https://doi.org/10.1029/2009GL038586, 2009.

Manney, G. L., Santee, M. L., Rex, M., Livesey, N. J., Pitts, M. C., Veefkind, P., Nash, E. R., Wohltmann, I., Lehmann, R., Froidevaux, L., Poole, L. R., Schoeberl, M. R., Haffner, D. P., Davies, J., Dorokhov, V., Gernandt, H., Johnson, B., Kivi, R., Kyrö, E., Larsen, N., Levelt, P. F., Makshtas, A., McElroy, C. T., Nakajima, H., Parrondo, M. C., Tarasick, D. W., von der Gathen, P., Walker, K. A., and Zinoviev, N. S.: Unprecedented Arctic Ozone Loss in 2011, Nature, 478, 469–475, 2011.

Manney, G. L., Lawrence, Z. D., Santee, M. L., Livesey, N. J., Lambert, A., and Pitts, M. C.: Polar processing in a split vortex: Arctic ozone loss in early winter 2012/2013, Atmos. Chem. Phys., 15, 4973–5029, 2015.

Manney, G. L., Hegglin, M. I., Lawrence, Z. D., Wargan, K., Millán, L. F., Schwartz, M. J., Santee, M. L., Lambert, A., Pawson, S., Knosp, B. W., Fuller, R. A., and Daffer, W. H.: Reanalysis comparisons of upper tropospheric/lower stratospheric jets and multiple tropopauses, Atmos. Chem. Phys., pp. 11 541–11 566, 2017.

McIntyre, M. E. and Palmer, T. N.: The “surf zone" in the stratosphere, J. Atmos. and Terr. Phys., 46, 825–849, 1984.

Molod, A., Takacs, L., Suarez, M., and Bacmeister, J.: Development of the GEOS-5 Atmospheric General Circulation Model: Evolution from MERRA to MERRA-2, Geosci. Model Dev., 8, 1339–1356, 2015.

Nash, E. R., Newman, P. A., Rosenfield, J. E., and Schoeberl, M. R.: An objective determination of the polar vortex using Ertel’s potential vorticity, J. Geophys. Res., 101, 9471–9478, 1996.

Nash, E. R., Strahan, S. E., Kramarova, N., Long, C. S., Pitts, M. C., Newman, P. A., Johnson, B., Santee, M. L., Petropavlovskikh, I., and Braathen, G. O.: Antarctic ozone hole [in "State of the Climate in 2015"]., Bull. Am. Meteor. Soc., 97, S168–S172, 2016.

Naujokat, B., Krüger, K., Matthes, K., Hoffmann, J., Kunze, M., and Labitzke, K.: The early major warming in December 2001 – exceptional?, Geophys. Res. Lett., 29, 2023, https://doi.org/10.1029/2002GL015316, 2002.

Newman, P. A., Kawa, S. R., and Nash, E. R.: On the size of the Antarctic ozone hole, Geophys. Res. Lett., 31, L21104, 2004.

Pawson, S., Krüger, K., Swinbank, R., Bailey, M., and O’Neill, A.: Intercomparison of two stratospheric analyses: Temperatures relevant to polar stratospheric cloud formation, J. Geophys. Res., 104, 2041–2050, 1999.

Pazmino, A. F., Godin-Beekmann, S., Ginzburg, M., Bekki, S., Hauchecorne, A., Piacentini, R. D., and Quel, E. J.: Impact of Antarctic polar vortex occurrences on total ozone and UVB radiation at southern Argentinean and Antarctic stations during 1997-2003 period, J. Geophys. Res., 110, D03103, https://doi.org/10.1029/2004JD005304, 2005.

Politis, D. N. and Romano, J. P.: The Stationary Bootstrap, Journal of the American Statistical Association, 89, 1303–1313, http://www.jstor.org/stable/2290993, 1994.

Polvani, L. M., Waugh, D. W., Correa, G. J., and Son, S.-W.: Stratospheric ozone depletion: The main driver of twentieth-century atmospheric circulation changes in the Southern Hemisphere, J. Clim., 24, 795–812, 2011.

Rex, M., Salawitch, R. J., Gathen, P., Harris, N. R., Chipperfield, M. P., and Naujokat, B.: Arctic ozone loss and climate change, Geophys. Res. Lett., 31, L04116, https://doi.org/10.1029/2003GL018844, 2004.

Rex, M., Salawitch, R. J., Deckelmann, H., von der Gathen, P., Harris, N. R. P., Chipperfield, M. P., Naujokat, B., Reimer, E., Allart, M., Andersen, S. B., Bevilacqua, R., Braathen, G. O., Claude, H., Davies, J., De Backer, H., Dier, H., Dorokhov, V., Fast, H., Gerding, M., Godin-Beekmann, S., Hoppel, K., Johnson, B., Kyrö, E., Litynska, Z., Moore, D., Nakane, H., Parrondo, M. C., Risley, A. D., Skrivankova, P., Stübi, R., Viatte, P., Yushkov, V., and Zerefos, C.: Arctic winter 2005: Implications for stratospheric ozone loss and climate change, Geophys. Res. Lett., 33, L23808, https://doi.org/10.1029/2006GL026731, 2006.

Rienecker, M. M. et al.: MERRA: NASA’s Modern-Era Retrospective Analysis for Research and Applications, J. Clim., 24, 3624–3648, 2011.

Riese, M., Ploeger, F., Rap, A., Vogel, B., Konopka, P., Dameris, M., and Forster, P.: Impact of uncertainties in atmospheric mixing on simulated UTLS composition and related radiative effects, Journal of Geophysical Research: Atmospheres, 117, n/a–n/a, https://doi.org/10.1029/2012JD017751, http://dx.doi.org/10.1029/2012JD017751, d16305, 2012.

Saha, S. et al.: The NCEP Climate Forecast System Reanalysis, Bull. Am. Meteor. Soc., 91, 1015–1057, 2010.

Saha, S. et al.: The NCEP Climate Forecast System Version 2, J. Clim., 27, 2185–2208, 2014.

Schoeberl, M. R. and Hartmann, D. L.: The dynamics of the stratospheric polar vortex and its relation to springtime ozone depletions, Science, 251, 46–52, 1991.

Schoeberl, M. R., Lait, L. R., Newman, P. A., and Rosenfield, J. E.: The structure of the polar vortex, J. Geophys. Res., 97, 7859–7882, 1992.

Schwartz, M. J., Lambert, A., Manney, G. L., Read, W. G., Livesey, N. J., Froidevaux, L., Ao, C. O., Bernath, P. F., Boone, C. D., Cofield, R. E., Daffer, W. H., Drouin, B. J., Fetzer, E. J., Fuller, R. A., Jarnot, R. F., Jiang, J. H., Jiang, Y. B., Knosp, B. W., Krüger, K., Li, J.-L. F., Mlynczak, M. G., Pawson, S., Russell, J. M., Santee, M. L., Snyder, W. V., Stek, P. C., Thurstans, R. P., Tompkins, A. M., Wagner, P. A., Walker, K. A., Waters, J. W., and Wu, D. L.: Validation of the Aura Microwave Limb SounderSic temperature and geopotential height measurements, J. Geophys. Res., 113, https://doi.org/10.1029/2007JD008783, http://dx.doi.org/10.1029/2007JD008783, 2008.

Simmons, A. J., Hortal, M., Kelly, G., McNally, A., Untch, A., and Uppala, S.: ECMWF analyses and forecasts of stratospheric winter polar vortex break-up: September 2002 in the southern hemisphere and related events, J. Atmos. Sci., 62, 668–689, 2005.

Simmons, A. J., Poli, P., Dee, D. P., Berrisfordand, P., Hersbach, H., Kobayashi, S., and Peubey, C.: Estimating low-frequency variability and trends in atmospheric temperature using ERA-Interim, Q. J. R. Meteorol. Soc., 140, 329–353, 2014.

Solomon, S.: Stratospheric ozone depletion: A review of concepts and history, Rev. Geophys., 37, 275–316, 1999.

Susskind, J., Rosenfield, J., and Reuter, D.: An accurate radiative transfer model for use in the direct physical inversion of HIRS and MSU temperature sounding data, J. Geophys. Res., 88, 8550–8568, 1983.

Takacs, L. L., Suárez, M. J., and Todling, R.: Maintaining atmospheric mass and water balance in reanalyses, Q. J. R. Meteorol. Soc., 142, 1565–1573, 2016.

Telford, P., Braesicke, P., Morgenstern, O., and Pyle, J.: Reassessment of causes of ozone column variability following the eruption of Mount Pinatubo using a nudged CCM, Atmos. Chem. Phys., 9, 4251–4260, 2009.

Tilmes, S., Müller, R., Engel, A., Rex, M., and Russel, J. M.: Chemical ozone loss in the Arctic and Antarctic stratosphere between 1992 and 2005, Geophys. Res. Lett., 33, https://doi.org/10.1029/2006GL026925, 2006.

Tomikawa, Y., Sato, K., Hirasawa, N., Tsutsumi, M., and Nakamura, T.: Balloon-borne observations of lower stratospheric water vapor at Syowa Station, Antarctica in 2013, Polar Science, 9, 345 – 353, https://doi.org/https://doi.org/10.1016/j.polar.2015.08.003, http://www.sciencedirect.com/science/article/pii/S1873965215300074, special Issue: The Asian Forum for Polar Sciences (AFOPS), 2015.

Waugh, D. W., Garfinkel, C. I., and Polvani, L. M.: Drivers of the Recent Tropical Expansion in the Southern Hemisphere: Changing SSTs or Ozone Depletion?, J. Clim., 28, 6581–6586, https://doi.org/10.1175/JCLI-D-15-0138.1, http://dx.doi.org/10.1175/JCLI-D-15-0138.1, 2015.

WMO: Scientific assessment of ozone depletion: 2006, Global Ozone Res. and Monit. Proj. Rep. 50, Geneva, Switzerland, 2007.

WMO: Scientific assessment of ozone depletion: 2014, Global Ozone Res. and Monit. Proj. Rep. 55, Geneva, Switzerland, 2014.

Wright, J. S., Fujiwara, M., Long, C., Anstey, J., Chabrillat, S., Compo, G. P., Dragani, R., Ebisuzaki, W., Harada, Y., Kobayashi, C., McCarty, W., Molod, A., Onogi, K., Pawson, S., Simmons, A., Tan, D., Wargan, K., Whitaker, J. S., and Zou, C.-Z.: SPARC Reanalysis Intercomparison Project (S-RIP) Final Report. Chapter 2: Description of the Reanalysis Systems, in preparation, to be published by SPARC in 2018.

Table 1. List of acronyms ~~of~~ for reanalysis assimilated observations and radiative transfer models

Acronym	Full Name
AIRS	Atmospheric Infrared Sounder
AMSR	Advanced Microwave Scanning Radiometer
AMSU	Advanced Microwave Sounding Unit
ATMS	Advanced Technology Microwave Sounder
ATOVS	Advanced Tiros Operational Vertical Sounder
COSMIC	Constellation Observing System for Meteorology, Ionosphere, and Climate
CrIS	Cross-track Infrared Sounder
CRTM	Community Radiative Transfer Model (radiative transfer model)
GLATOVS	Goddard Laboratory for Atmospheres TOVS (radiative transfer model)
GMS	Geostationary Meteorological Satellite
GOES	Geostationary Operational Environmental Satellite
GPS-RO	Global Positioning System Radio Occultation
HIRS	High resolution Infrared Radiation Sounder
IASI	Infrared Atmospheric Sounding Interferometer
MHS	Microwave Humidity Sounder
MLS	(Aura) Microwave Limb Sounder
MSU	Microwave Sounding Unit
MTSAT	Multi-functional Transport Satellite
RTTOV	Radiative Transfer for TOVS (radiative transfer model)
SSM/I	Special Sensor Microwave Imager
SSMIS	Special Sensor Microwave Imager Sounder
SSU	Stratospheric Sounding Unit
TMI	Tropical Rainfall Measuring Mission Microwave Imager
TOVS	Tiros Operational Vertical Sounder
VTPR	Vertical Temperature Profile Radiometer

~~Schematic example of steps to go from daily differences to yearly time series. The top panel shows one year (1995) of daily minimum temperatures from MERRA-2; purples/blues (reds) represent low (high) values on the color bar. These values are subtracted from the corresponding ones for each of the reanalyses, yielding fields such as those shown (for ERA-Interim – MERRA-2) in the lower left; here positive (ERA-Interim greater than MERRA-2) differences are shown in reds, negative (ERA-Interim less than MERRA-2) differences in blues. These values are averaged over the period indicated by the black vertical lines to get a number for each level (numbers to the right of difference plot), and those are plotted as a stacked array of squares for each year (lower right).~~

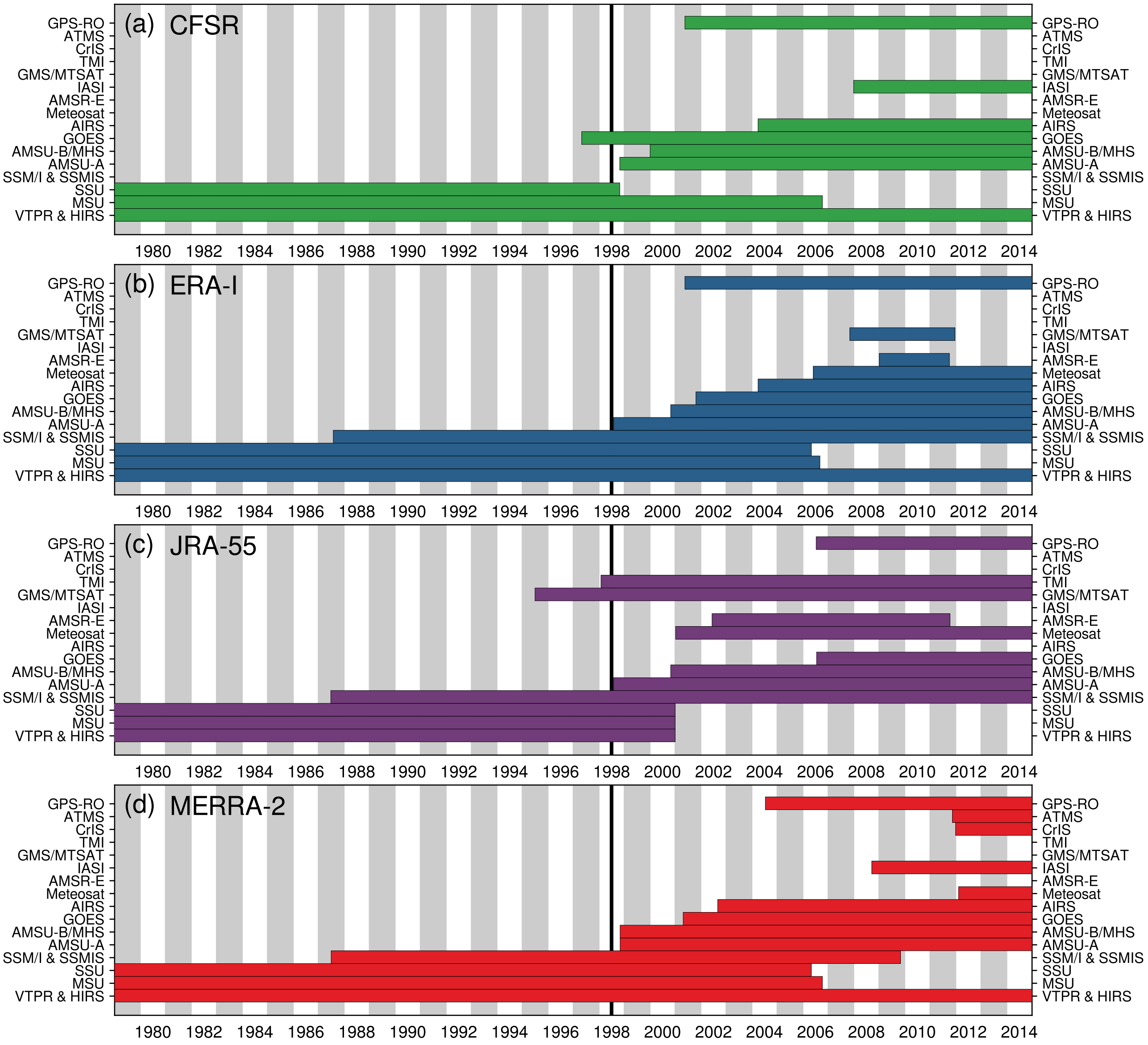


Figure 1. Time line for operational satellite instrument inputs to the reanalyses used herein: panels (a) through (~~e~~d) show CFSR/CFSv2, ERA-I, JRA-55, ~~MERRA,~~ and MERRA-2, respectively. Table 1 gives a list of the acronyms used here. Within the constraint of putting them in the same order for each reanalysis, the input datasets are stacked in approximately chronological order, with earliest on the bottom and latest on the top. The black vertical line is at mid-1998, near the time of the TOVS to ATOVS transition (see text). See Fujiwara (Fujiwara et al., 2017) for a similar time line but organized per instrument.

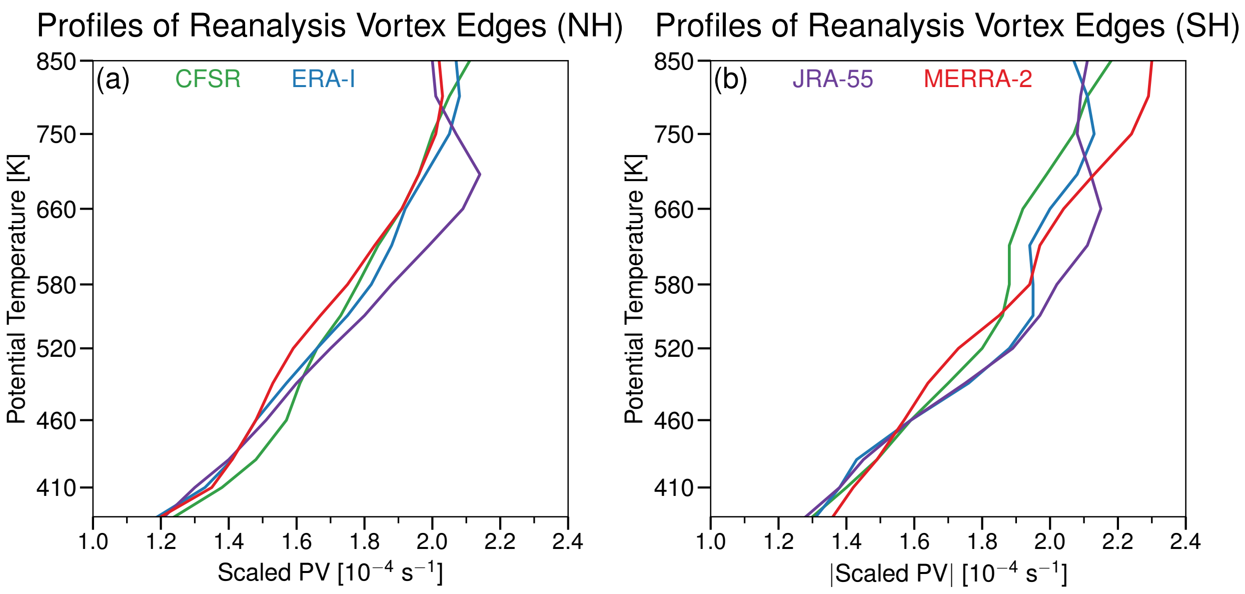


Figure 2. ~~Schematic illustration of how "agreement" among reanalyses is assessed. The cyan and magenta lines show (top) the average~~ Potential temperature profiles of the ~~differences between two hypothetical reanalyses and another~~ reanalysis ~~used~~ vortex edges determined from climatological maximum PV gradients expressed as scaled PV for the (a~~reference~~) NH, and the (~~bottom~~b) ~~the standard deviation of the differences that were averaged. See text for description of how these indicate changes in agreement~~SH.

~~Total number of days with temperatures below T_{ice} in the SH for each year summed over the lower stratosphere (from 121.1 to 31.6hPa) (a and c), and the sensitivity ranges of each reanalysis calculated using ±1K offsets from T_{ice} (b and d). The colored bars show the range of values obtained from the tests of sensitivity to the PSC threshold temperature used (see Section 2.2.2), while the black dots show the value for the "central" threshold temperature. The days are counted from April through the following February. For each year, the reanalyses are ordered from smallest central value on the left to largest central value on the right; this order is also given as a text string at the top of the column for each year. The numbers at the bottom of each year's column indicate the difference in days between the largest and smallest central values for the year (i.e., between the rightmost and leftmost black dots). In the range panels (b and d), the range about the central value (black dots in a and c) is shown for each reanalysis. Green, blue, purple, pink, and red indicate CFSR, ERA-I, JRA-55, MERRA, and MERRA-2, respectively.~~

~~As in Figure **??** but for the NH and T_{NAT}. The days are counted from October through May.~~

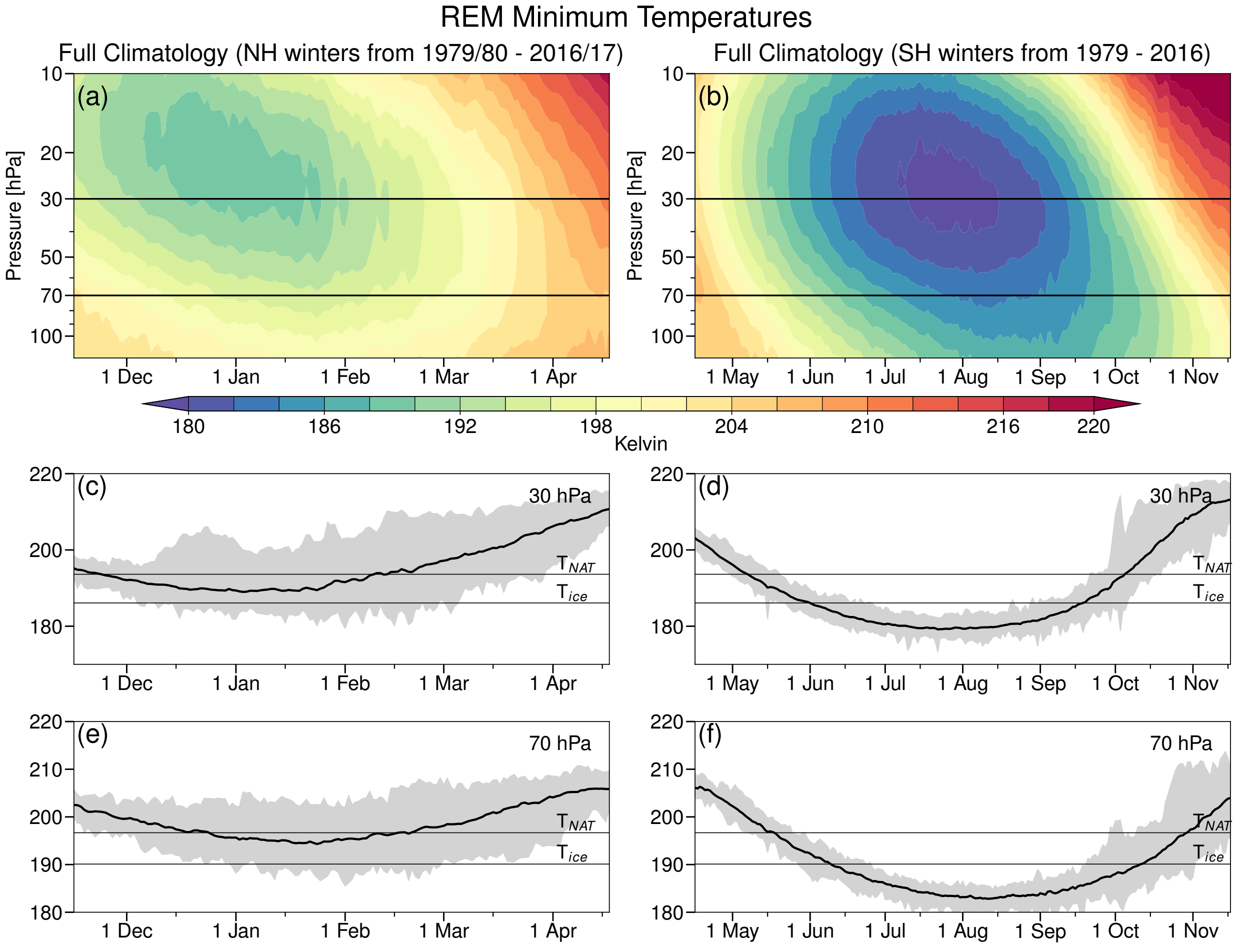


Figure 3. Time series of reanalysis ensemble mean (REM) (a, c, e) Arctic and (b, d, f) Antarctic climatological (1979/1980 through 2016/2017 in the NH, and 1979 through 2016 in the SH) minimum high latitude (poleward of 40° latitude) temperatures~~in the MERRA-2 reanalysis~~; (a) and (b) show contour plots (blues/purples represent low temperatures and reds high temperatures)~~, with line plots shown at~~. The horizontal black lines mark the 30 hPa and 70 hPa pressure levels for which line plots are shown in (c) and (d) and ~~at 70hPa~~ in (e) and (f), respectively. The shading in the ~~lines~~ line plots shows the range of REM values on each date, and the black line the average values. Note that the time period shown is longer in the SH than in the NH. The same color range is used for each hemisphere.

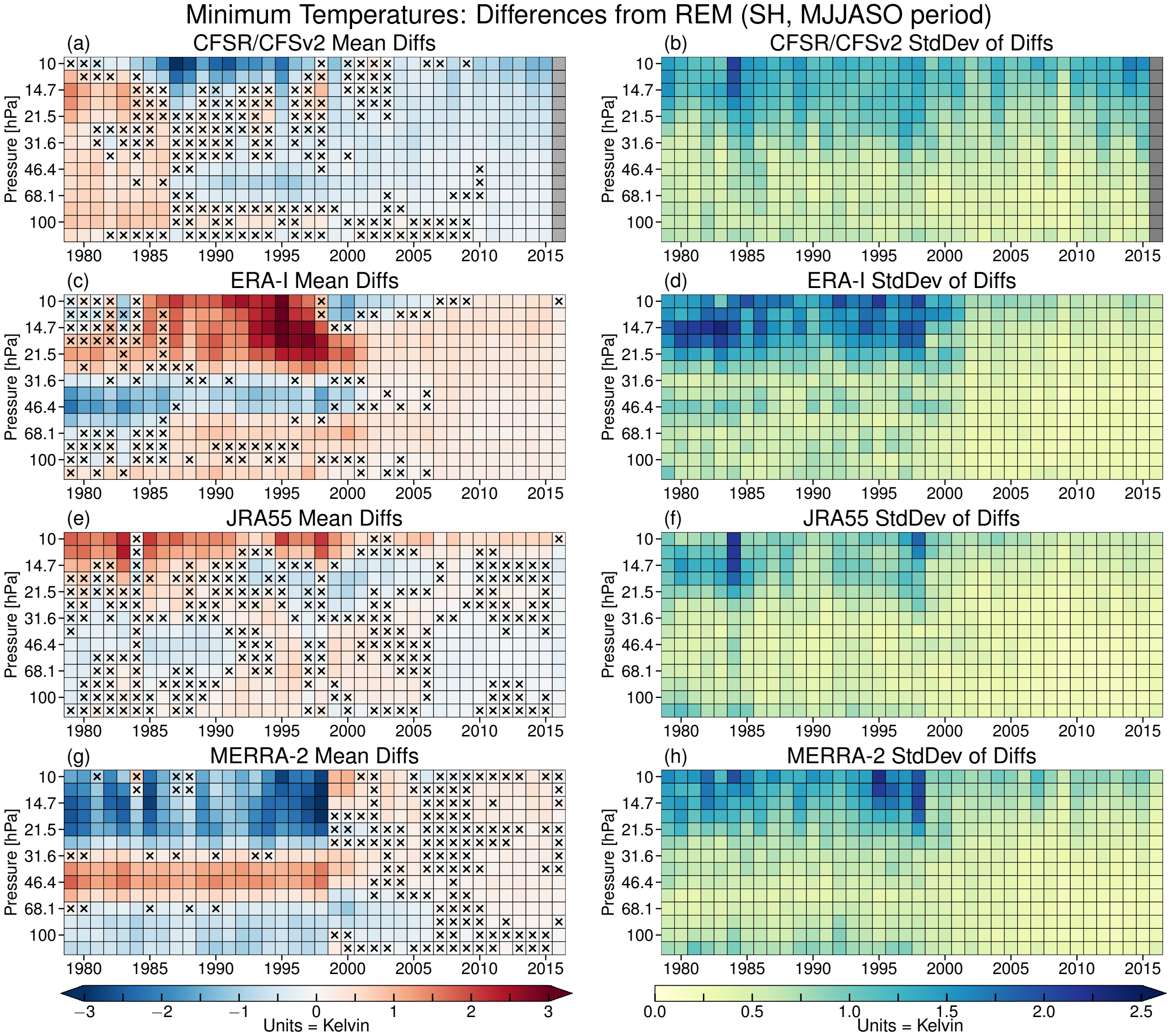


Figure 4. SH winter season (MJJASO) (a, c, e, g) averages and (b, d, f, h) standard deviations of minimum temperature differences for each reanalysis from ~~MERRA-2~~ the reanalysis ensemble mean (REM) as a function of year and pressure for the 1979 through 2015 winters, concatenated from the individual years into pixel plots as described in the text~~and Figure ??~~. Columns of grey pixels indicate years with no data. Pixels with x symbols inside indicate years and levels where the differences from the REM are insignificant according to our bootstrapping analysis (see section 2.2.4). Blues in the average difference panels show negative values (reanalysis less than ~~MERRA-2~~the REM) and reds positive values (reanalysis greater than ~~MERRA-2~~the REM); in the standard deviation panels, yellows/deep blues represent low/high standard deviations of the reanalysis differences, respectively.

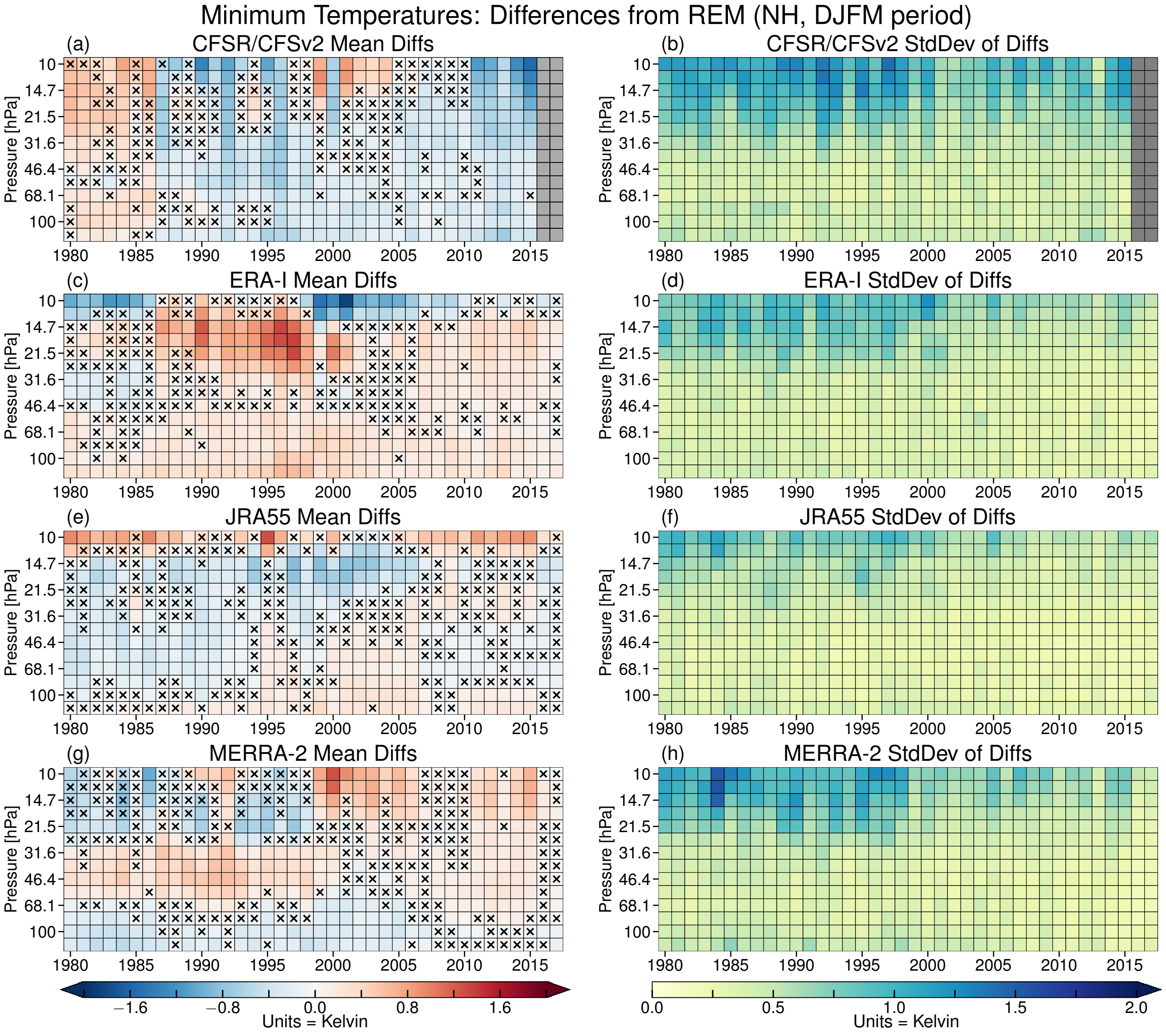


Figure 5. As in Figure 4 but for the NH winter seasons (DJFM) for 1979/1980 through 2015/2016. Note that different color ranges are used for the NH shown here than in Figure 4 for the SH.

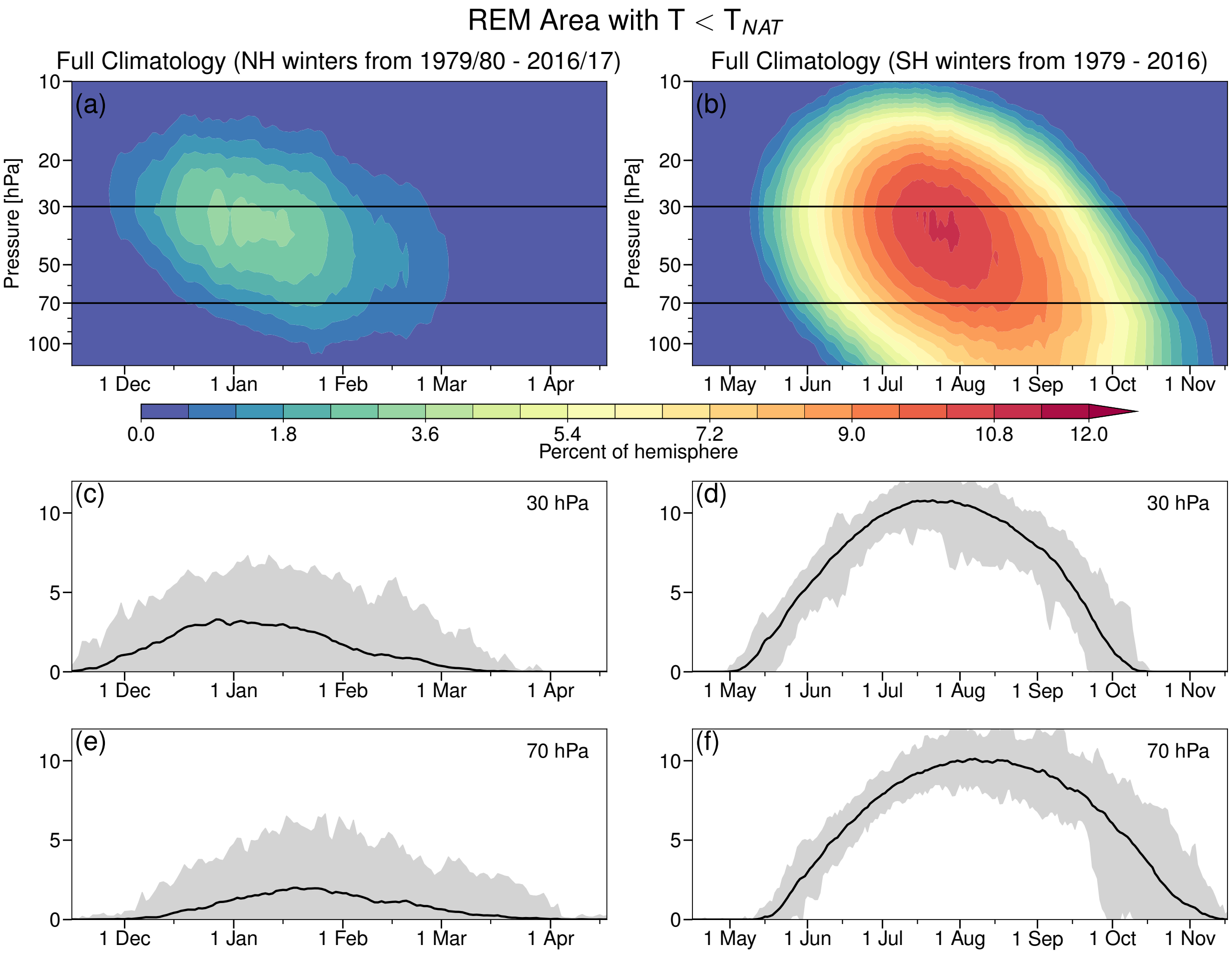


Figure 6. As in Figure 3, but for area with temperatures below the NAT PSC threshold.

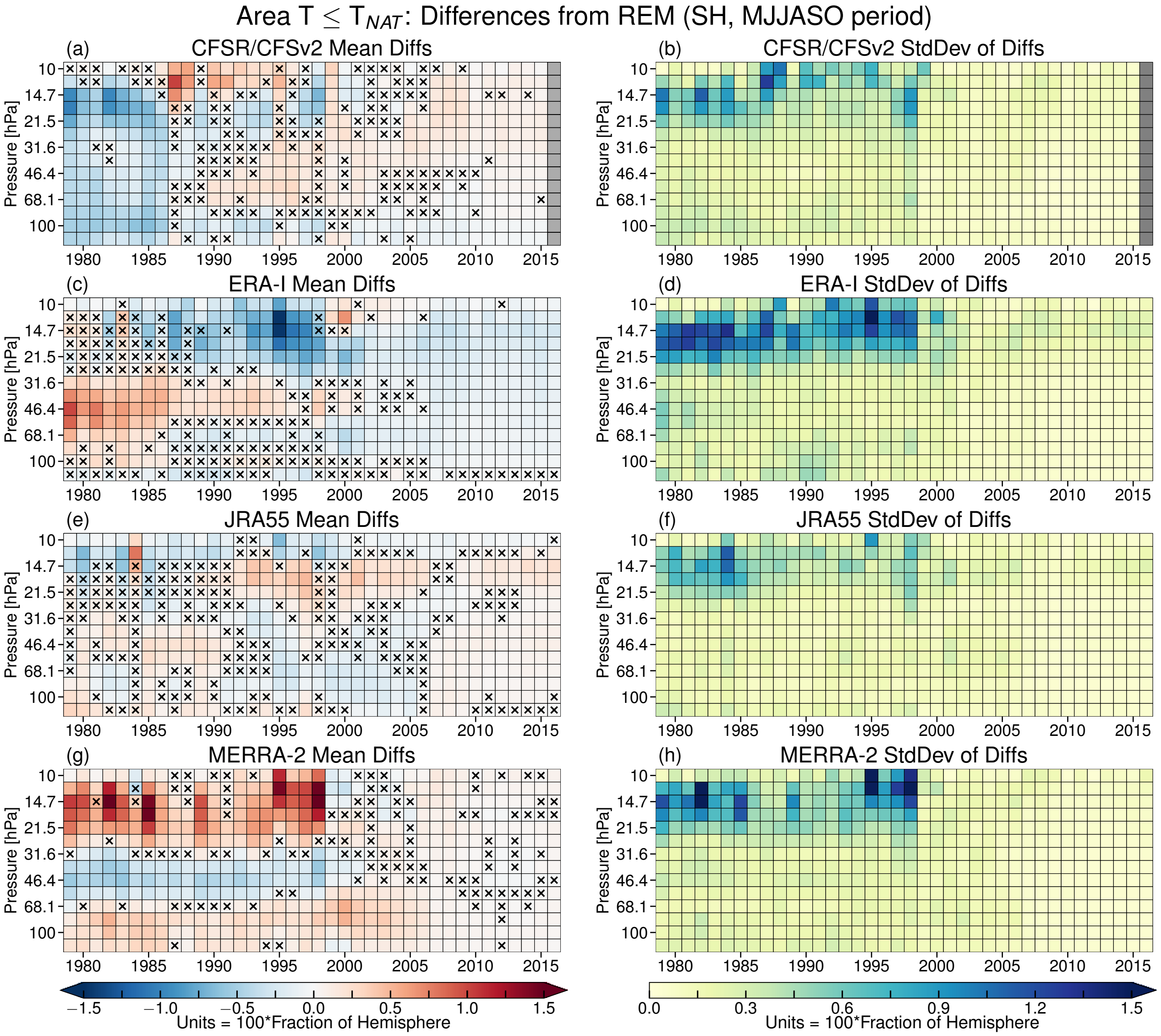


Figure 7. As in Figure 4 but for area with temperatures below the NAT PSC threshold in the SH.

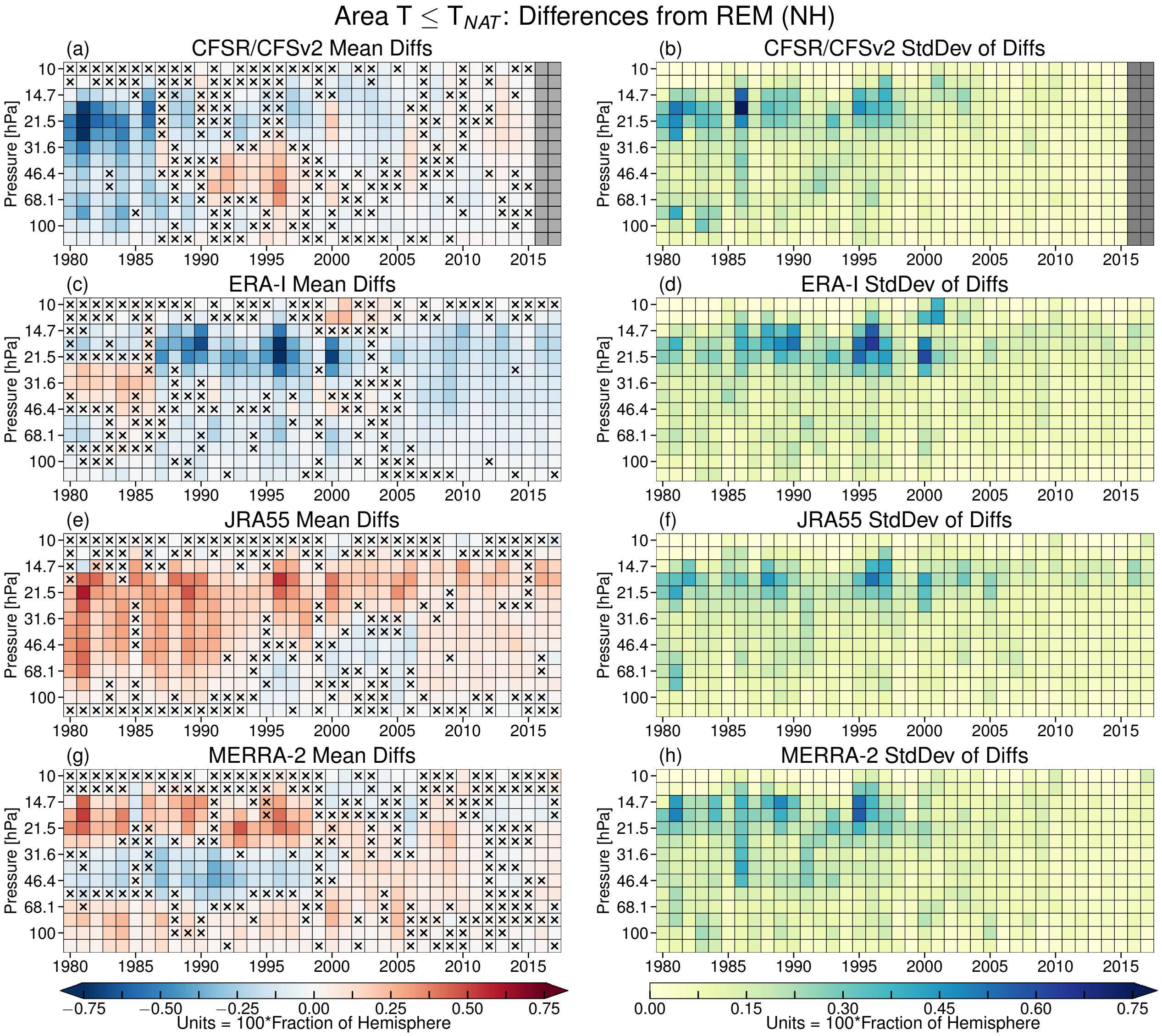


Figure 8. As in Figure 7, but for the NH. See text for explanation of date ranges used for the calculations.

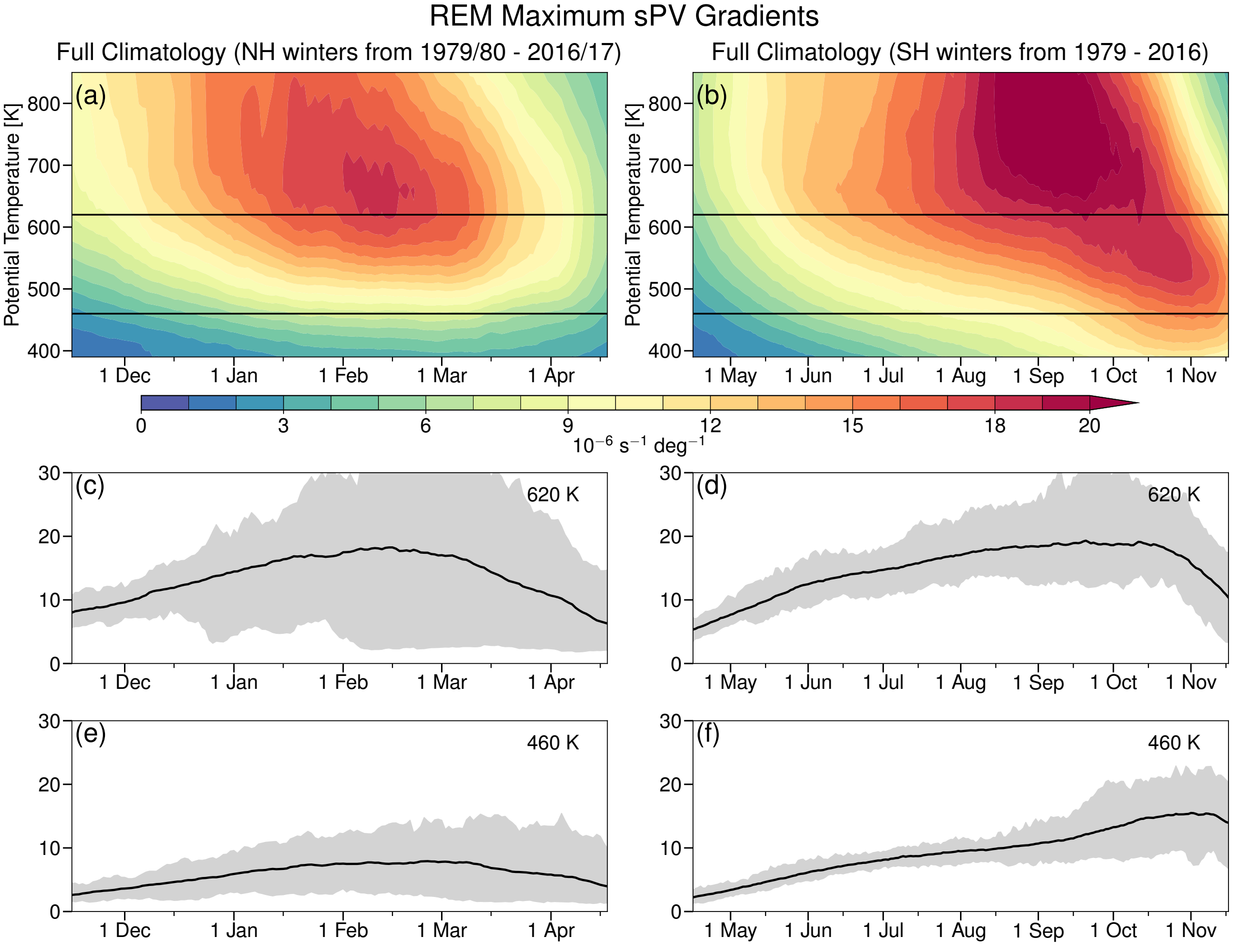


Figure 9. As in Figure 3, but for maximum sPV gradients.

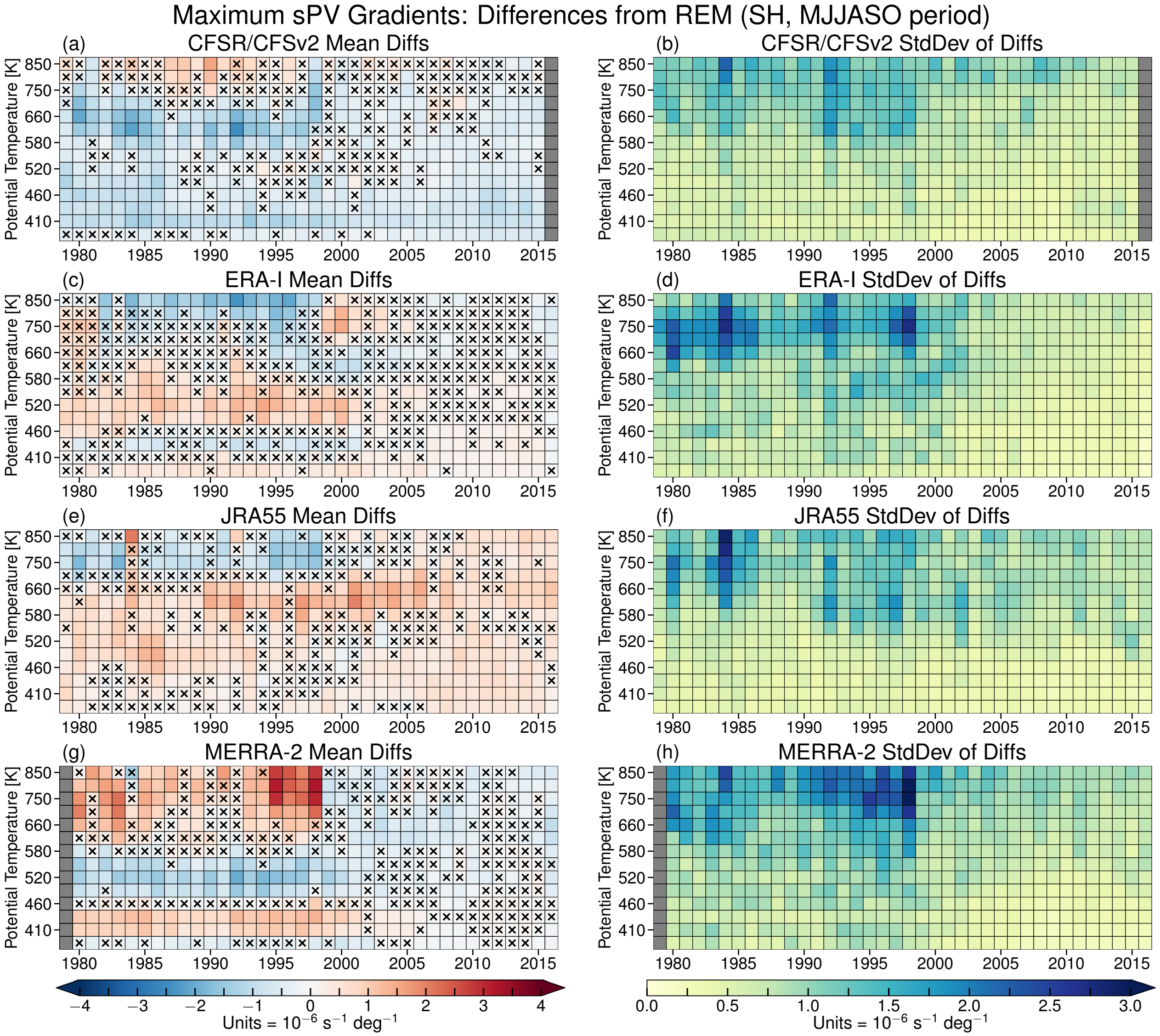


Figure 10. As in Figure 4, but for maximum sPV gradients.

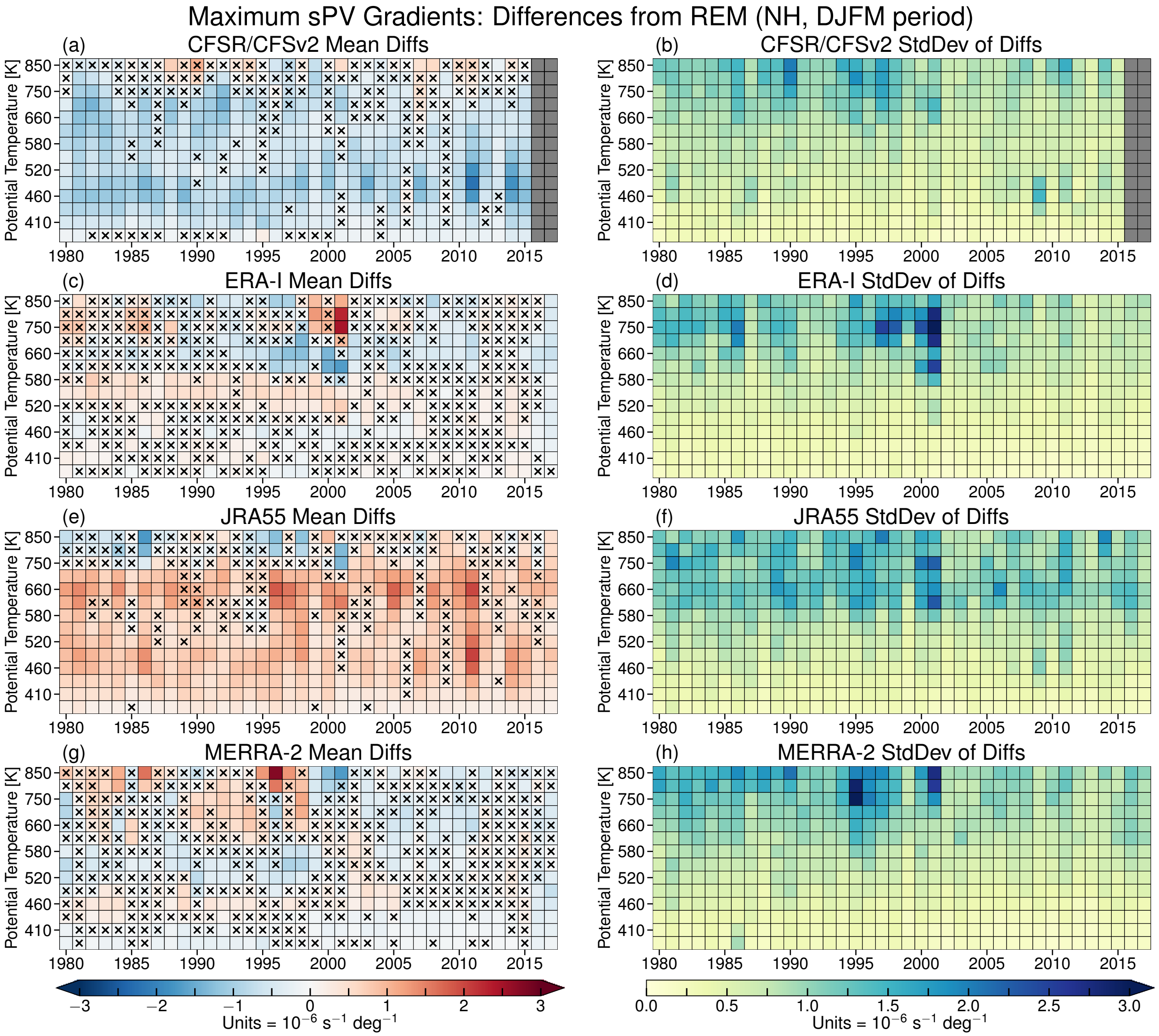


Figure 11. As in Figure 10, but for the NH.

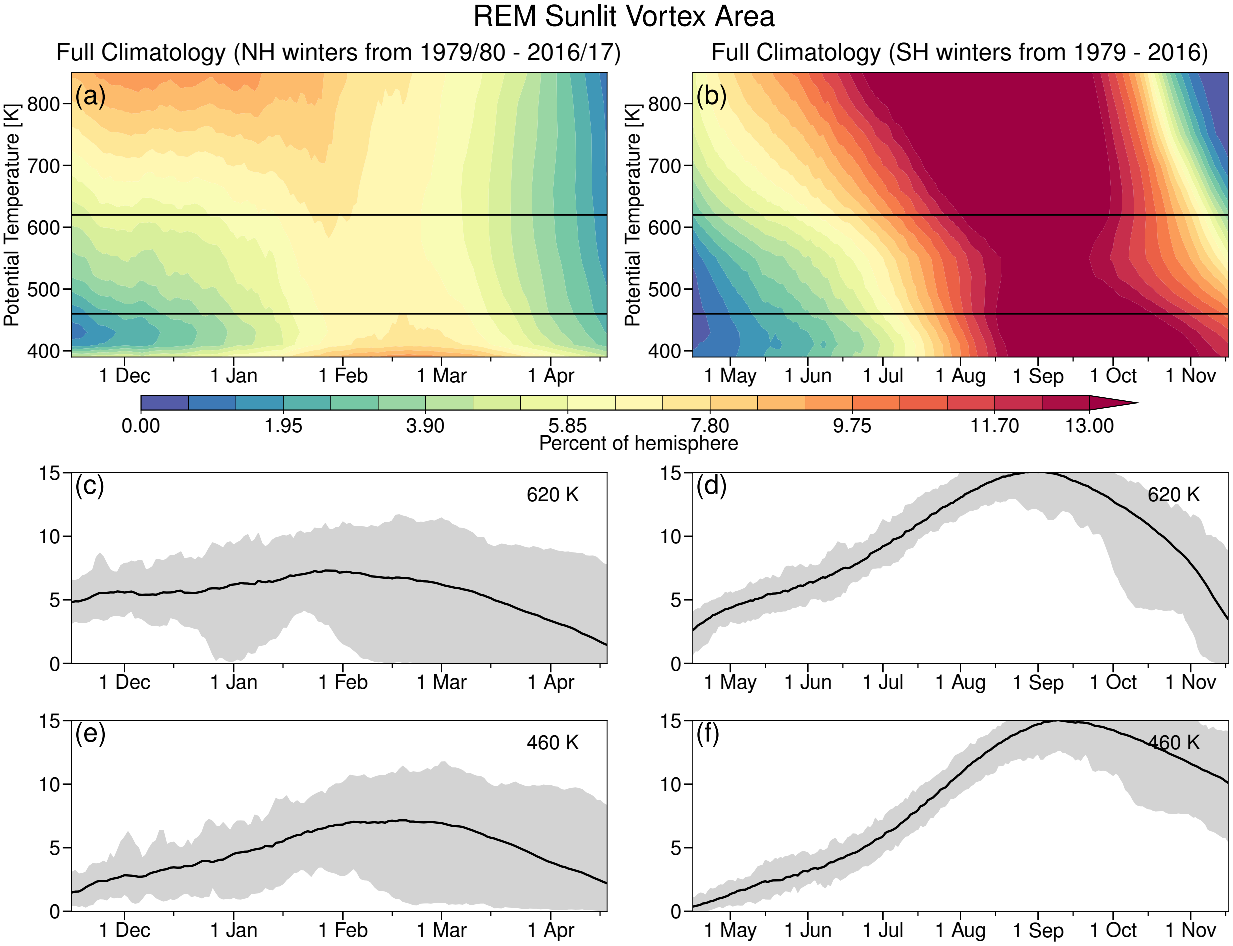


Figure 12. As in Figure 9, but for sunlit vortex area.

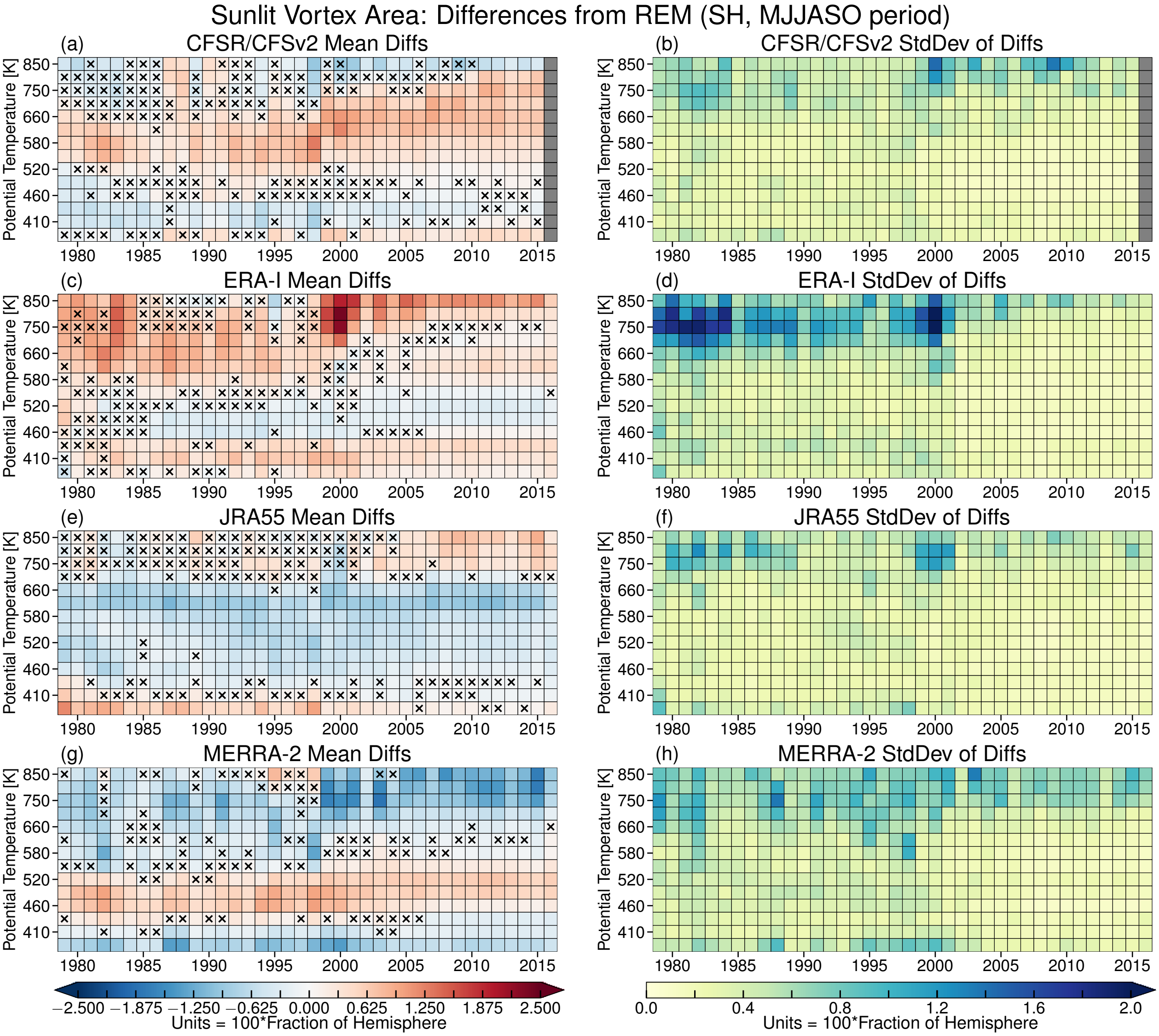


Figure 13. As in Figure 4, but for SH sunlit vortex area.

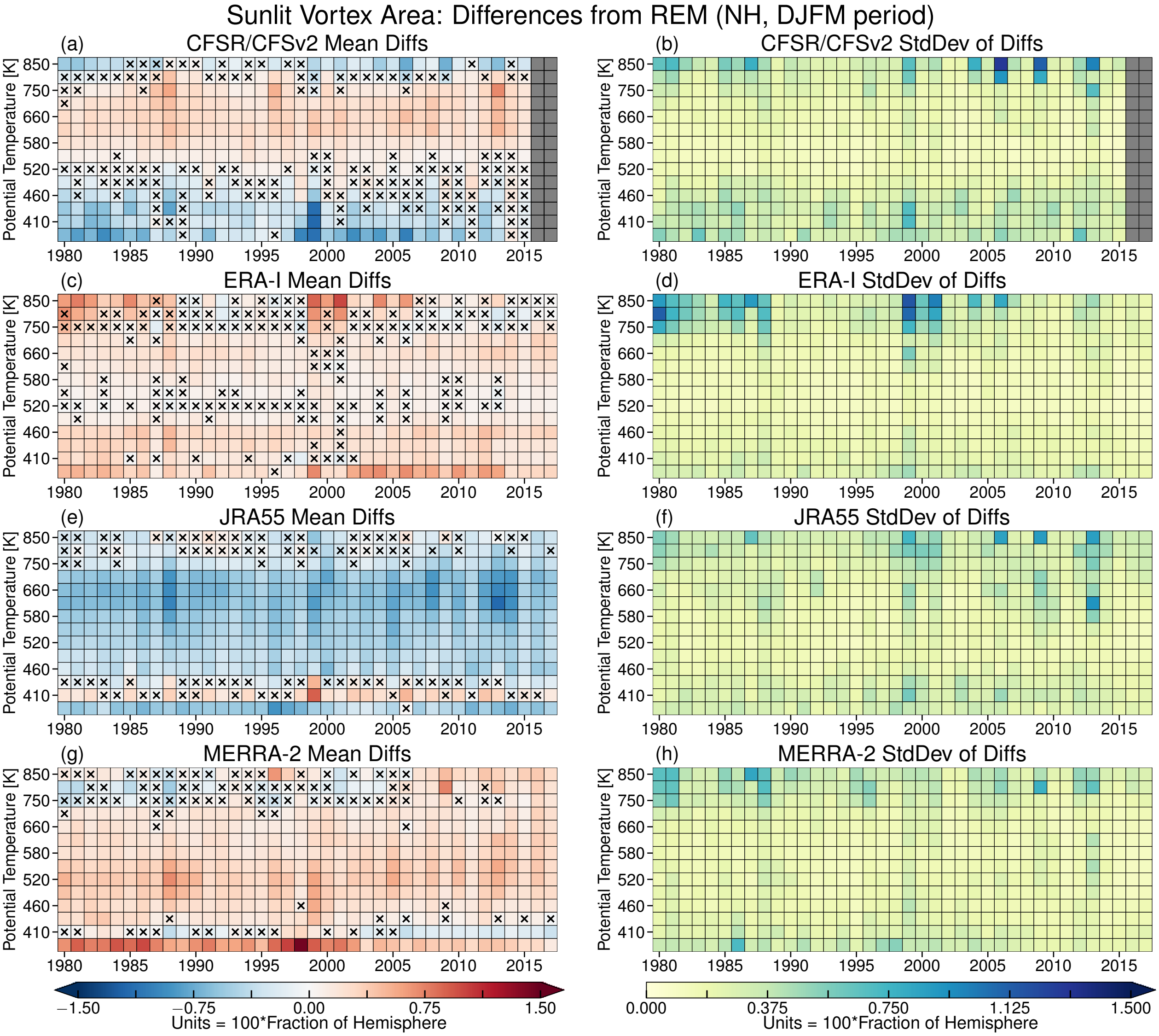


Figure 14. As in Figure 13, but for the NH.

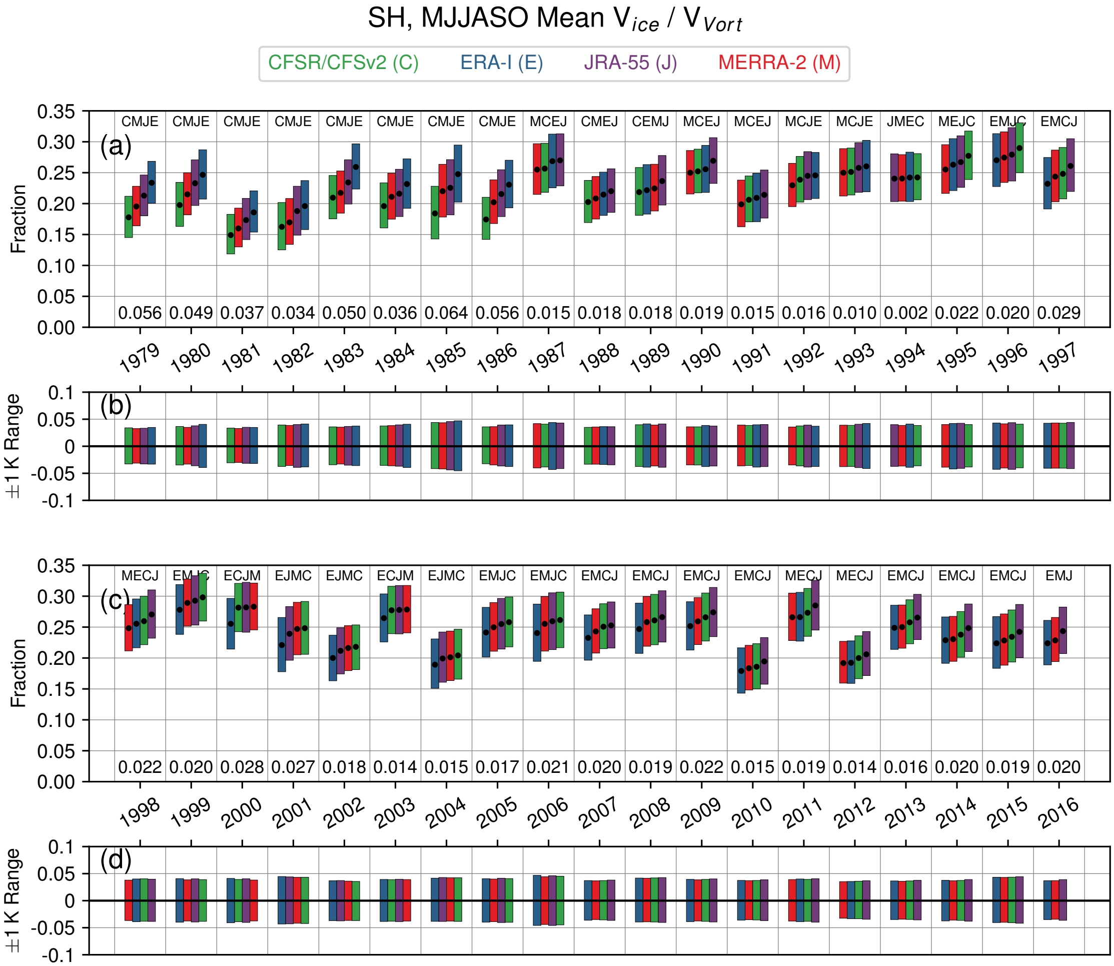


Figure 15. Winter ~~mean~~ means of the fraction of vortex volume between the 390 and 580 K isentropic surfaces with temperatures below T_{ice} in the SH (a and c), and range of values obtained for the ±1 K sensitivity tests (b and d). The colored bars show the range of values obtained from the tests of sensitivity to the PSC threshold temperature used (see Section 2.2.2), while the black dots show the value for the "central" threshold temperature. The winter mean is calculated over the full MJJASO period. ~~The layout~~ For each year, the reanalyses are ordered from smallest central value on the left to largest central value on the right; this order is also given as ~~in Figure ??, with~~ a text string at the top of the column for each year. The numbers at the bottom of ~~(a) and (c) being~~ each year's column indicate the ~~range of~~ difference in winter mean fraction between the largest and smallest central values for the winter season (~~that is~~i.e., between the rightmost ~~minus~~ and leftmost black dots). In the range panels (b and d), the range about the central value (black dots in a and c) is shown for each reanalysis. Green, blue, purple, ~~pink,~~ and red indicate CFSR/CFSv2, ERA-I, JRA-55, ~~MERRA,~~ and MERRA-2, respectively.

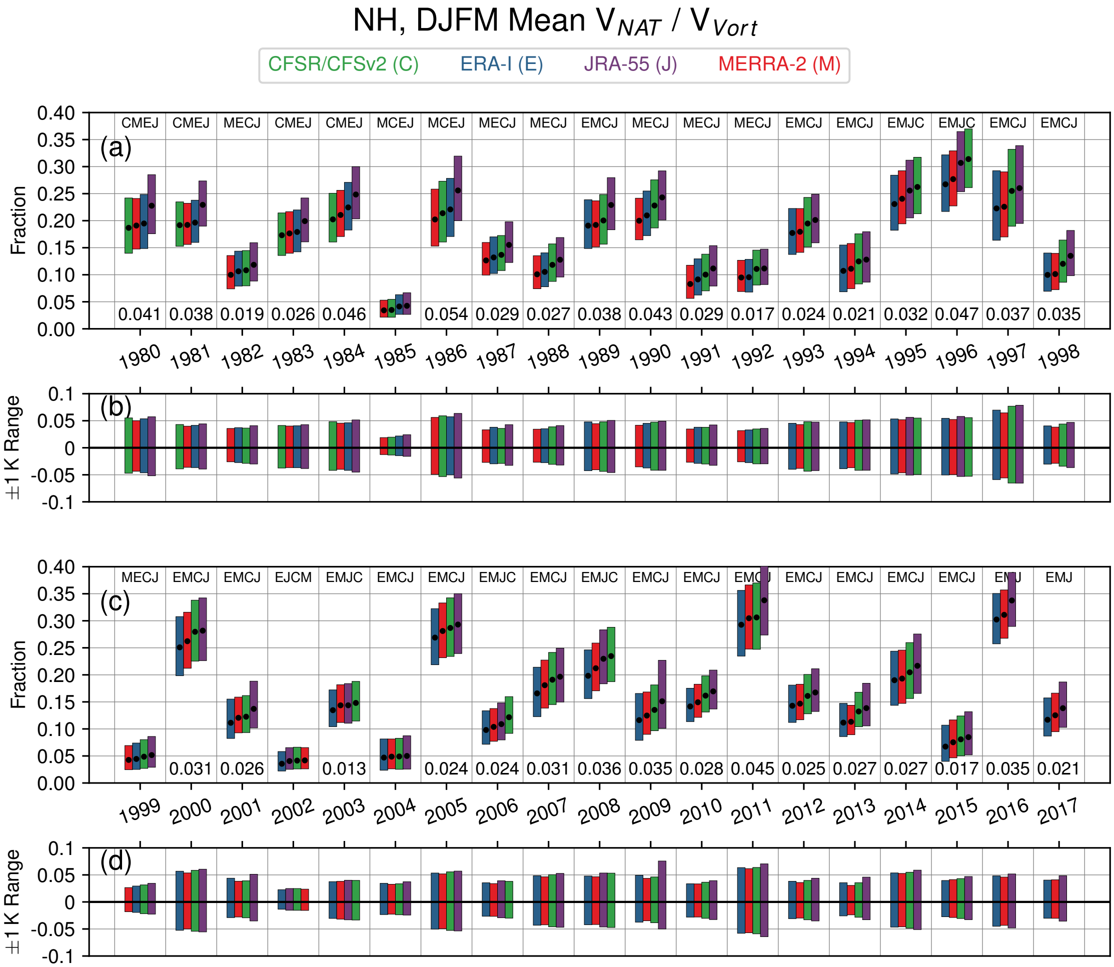


Figure 16. As in Figure 15 but for the NH and temperatures below T_{NAT}

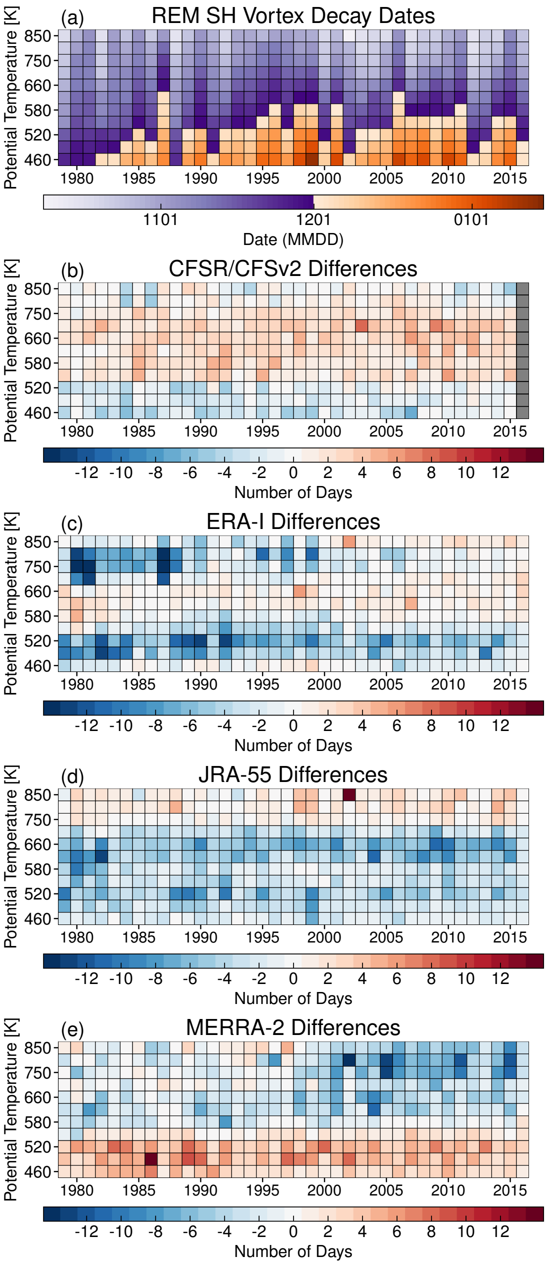


Figure 17. Pixel plots of (a) vortex decay dates (see text for the definition) ~~in~~ based on the ~~MERRA-2~~ reanalysis ensemble mean (REM) of vortex area, and (b through e) the difference between the vortex decay dates in each of the ~~other~~ reanalyses ~~and MERRA-2~~ from the REM (as reanalysis − ~~MERRA-2~~REM). The color bar ranges are restricted to distinguish differences of a few days; differences ~~that~~ whose magnitude greatly ~~exceed~~ exceeds the range (by more than 7 days, thus differences with magnitude greater than 21 days) are marked with a white X.

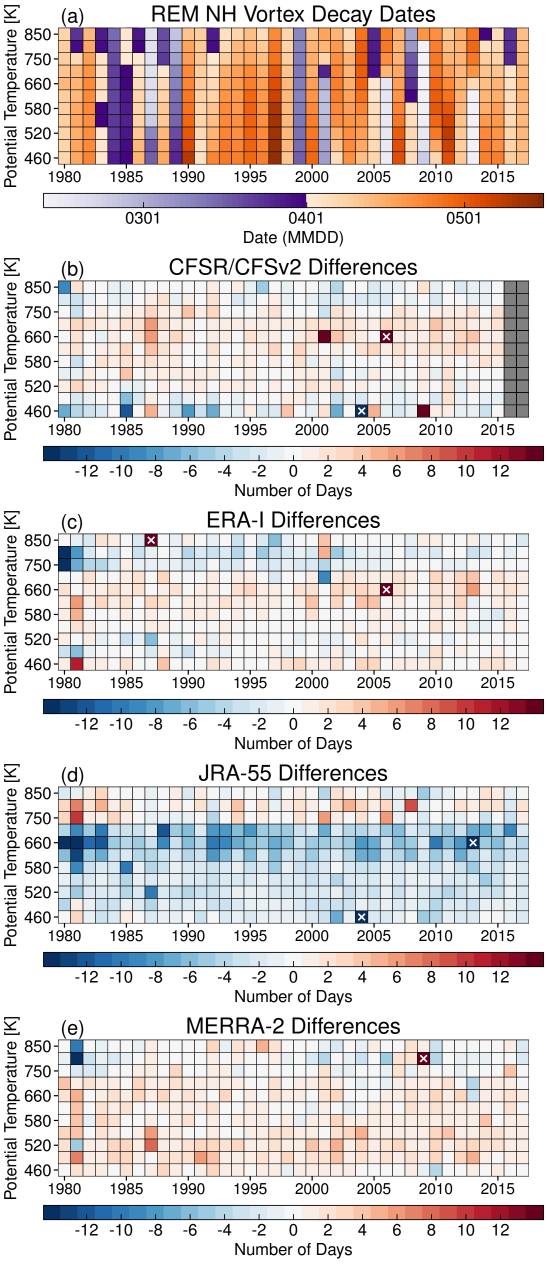


Figure 18. As in Figure 17 but for the NH.

Table A1. HNO_3 and H_2O Values for PSC Threshold Calculation. Pressure values are rounded to nearest integer.

pressure (hPa)	HNO_3 (ppbv)	H_2O (ppmv)
147	3.0	4.0
100	3.7	4.0
68	6.1	4.8
46	9.8	5.0
32	12.0	5.5
22	11.8	5.8
15	10.0	5.9
10	6.8	6.0

Table A2. Pressure levels, approximately corresponding potential temperature levels, and the NAT and ice PSC thresholds calculated from Table A1. Pressure values are rounded to the nearest integer; NAT and ice PSC thresholds are rounded to the nearest tenth of a Kelvin.

pressure (hPa)	potential temperature (K)	NAT PSC threshold (K)	ice PSC threshold (K)
121	390	198.2	192.4
100	410	197.3	191.2
83	430	197.1	190.7
68	460	196.6	190.0
56	490	196.0	189.0
46	520	195.3	188.0
38	550	194.6	187.2
32	580	193.9	186.4
26	620	192.9	185.5
22	660	192.0	184.5
18	700	190.9	183.5
15	750	189.8	182.5
12	800	188.6	181.5
10	850	187.4	180.5