Referee comment on "Reanalysis intercomparisons of stratospheric polar processing diagnostics" by Lawrence et al., ACPD, 2018.

Simon Chabrillat (Royal Belgian Institute for Space Aeronomy, BIRA-IASB)

April 15, 2018

1 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.

2 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.

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.

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 T_{min} 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.

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*~~ *reveal****s 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).

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*".
- P.9, line 2: please provide the climatological profiles of H_2O and HNO_3 used to determine T_{ICE} and T_{NAT} (with original reference if available).
- 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).

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).

3 Minor comments and corrections

- Abstract: long enough that it would be useful to split it into two or three paragraphs.
- 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.
- 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 $T < T_{NAT}$.
- 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.*"
- P.2, line 23: "*the conversion of* ***chlorinated species*** *into forms...*". Please also mention brominated species.
- P.2, line 29: Please define "active chlorine" (i.e. Cl and ClO)
- 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...*"
- P.4, line9: CFSR is not considered as a "full-input reanalysis"? Why?
- P.4, line 26: "*... during much of which ...*" - please re-phrase
- 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?).
- 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).
- 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,...*".
- P.5, line 19: define acronym "GMAO" or maybe drop it (anyway GMAO is the only division delivering atmospheric reanalyses at NASA).
- 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).

- P.6, line 8: "*...but a much ~~older~~* ***earlier*** *version than ~~that used~~ in MERRA-2*".
- P.6, line 15: Please replace this URL by 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*".
- 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).
- Section 2.2.1: Consider citing also Manney et al. (JGR, 2007) which provided an excellent overview about the Derived Meteorological Products used here.
- 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 eep, line 23: replace "*To demonstrate...*" by "*To illustrate...*".
- P.11, lines 19–21: AIRS was introduced on September 2002 in MERRA and MERRA-25 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.
- P.12, line 30: "*...(as was the case in the SH)* ***they*** *remain larger in CFSR...*"
- P.12, line 35: "*...a clear decrease in them is not evident*". Please re-phrase.
- 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 T_{min} 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.
- P.14 line 29: remove words "*... show that the variances of the differences...*"
- P.15 line 6: see corresponding major comment (2.3)
- P.15 line 37: "*... (indicating* ***that*** *the****se*** *reanalyses have~~ing~~ ~~higher~~* ***larger*** *SVA* ***than MERRA-2****)...*"
- 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.
- 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).

- 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).

- 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?

- P.17 line 21: Why is the important diagnostic V_{PSC}/V_{vort} and not V_{PSC} itself? If possible, explain this in one additional sentence.

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

- 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.

- 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).

- 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.

- 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.*"

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

- 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.*

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

- 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.

- 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.