We thank the reviewer for her/his comments. Below are our responses in blue. The biggest change is that the update version includes new figures in the appendix showing root mean square (RMS) differences of the parameters studied to get an idea of the day-to-day variability. Brief text explaining these RMS differences was added throughout the manuscript.

Millán and colleagues study the differences of potential vorticity (PV) and PV-based diagnostics in four modern reanalyses (ERA-Interim, MERRA-2, JRA-55 and CFSR/CFSv2). The discussion centers around (i) the calculation of PV and the differences arising in this task in the various reanalyses, (ii) the impact of the data assimilation on PV in each reanalysis product, (iii) seasonal and annual mean variability of sPV between the various reanalyses, as well as of PV-based diagnostics such as (iv) equivalent latitude, (v) dynamic tropopause and (vi) polar vortex characterization. The major finding is that PV agrees well between the various data sets on the time scales studied in this work. The authors also highlight the situation where more caution is necessary when working with PV. Some differences between the various data sets arise in particular for (i) equivalent latitude calculations at low latitudes or high altitudes, (ii) the dynamic tropopause in regions of jetstreams and of strong topography, as well as (iii) during the formation and demise of the polar vortex.

This work is intended as part of the S-RIP special issue where it perfectly fits. Such a comparison of PV from different reanalysis data sets has not been completed yet, although PV from reanalysis is a widely used diagnostic to analyze transport and dynamics in the troposphere and stratosphere. The questions asked in the paper are clear. The analysis is very convincing; data and methods are well described. The figures are well structured and clear to understand. The conclusions are based on the analysis. It is easy to follow the thoughts of the authors and as a reader I have the feeling that the authors really know what they are talking about. In general, I think this study will be of great value to users of reanalysis data and as such I would support publication of this study in ACP in the S-RIP special issue. I have some rather minor comments listed below, which the authors might consider for a revised version.

Comments:

• P3, L11: To my knowledge, ERA-I provides relative vorticity, see eg. the ERAInterim data catalogue: https://apps.ecmwf.int/archive-catalogue/?stream=oper&levtype=ml&expver=1&month=jan&year=1979&type=an&class=ei

Great catch, thanks! The sentence will now read: "CFSR/CFSv2 provides absolute vorticity while ERA-Interim provides relative vorticity, hence, for …"

• Equation 1: maybe it is worth mentioning that this is the synoptic approximation of PV. We added: "By using the provided or derived vorticity fields, we make use of the synoptic approximation to calculate PV, which assumes that ς_θ+f approximately equals the absolute vorticity and that horizontal gradients of potential temperature are small."

• Equation 2: I think, since not everybody might be familiar with sPV, some readers would benefit from a comparison of sPV and PV, maybe shown here for one reanalysis but for different averaging times, e.g., a snapshot, monthly, or yearly mean.

We will include the following figure:

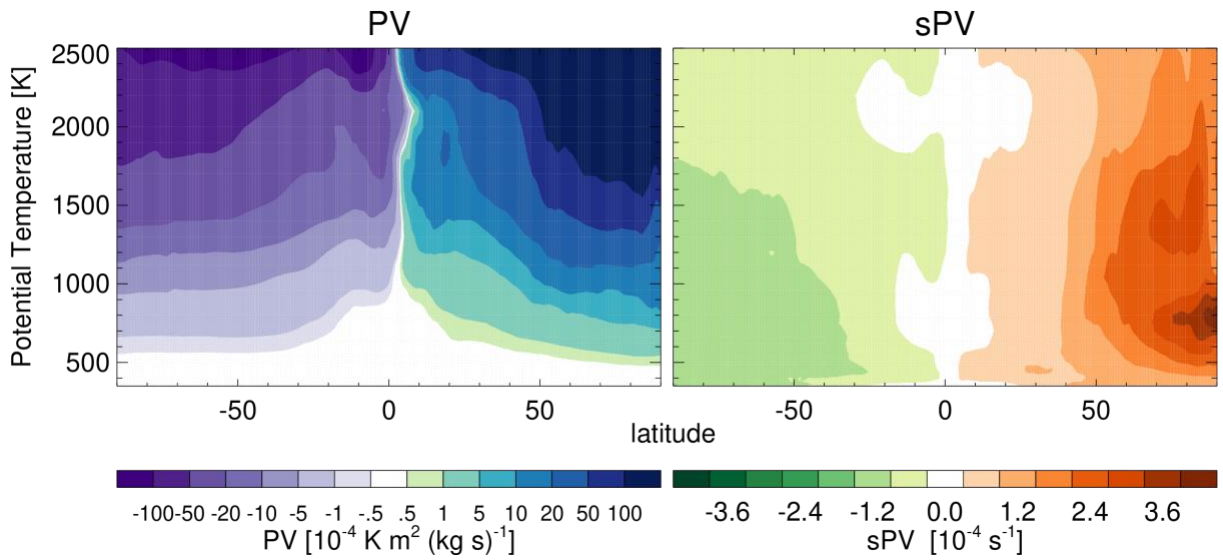


Figure 1: January 1st 2005 PV (left) and sPV (right). Note that sPV has similar order of magnitude values throughout the stratosphere as opposed to PV (for which color bar is non-linear).

The text about sPV will changed to: "This scaling is performed to provide fields with a similar order of magnitude throughout the stratosphere as opposed to PV (which increases approximately exponentially with increasing θ**, as shown in Figure 1**)."

• P3, l26: Could you mention the potential temperature range here. We added in brackets "(330, 340, 360, 380, 400, 420, 440, 460, 480, 500, 520, 540, 560, 580, 600, 620, 660, 700, 750, 800, 850, 900, 960, 1040, 1120, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500). "

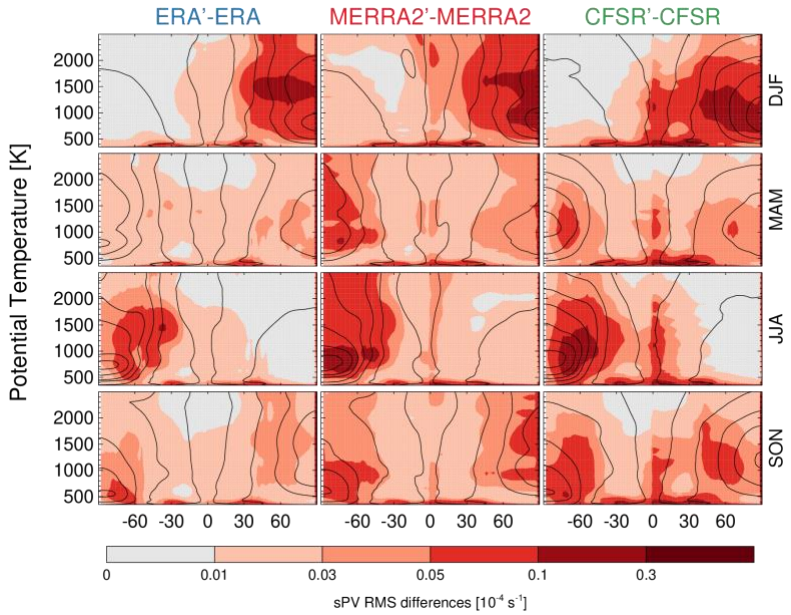
• P4, l34: Could you say something about the cause of the low bias of sPV? Is this an effect of vertical grid spacing or model physics (e.g. GW drag)?

We do not really know, we did add that in the next figure (the one discussing the discontinuities) that the constant bias at 2500K is likely related to the low lid height for CFSR/CFsV2. This was also mention in the EqL comparison.

• Sec. 4: When there is a 4.1, there should be a 4.2 as well. I would suggest to find a subhead for the first paragraphs of this section.
Subsection 4.2 was converted into a new section with the title "Variations due to differing calculation methods"

• Figure 4 and related discussion: Is Fig. 4 based on the temporal and zonal mean of the entire data set? Can you say something about the order of the magnitude of the differences, if shorter time periods are considered (monthly means or even shorter).

We will include the following figure in the appendix



Caption: Seasonal root-mean-square (RMS) daily (1980-2014) sPV differences between the sPV from each reanalysis provided vorticity or PV and the sPV computed from that reanalyses' horizontal wind, pressure, and temperature fields. Overlaid contours show each reanalyses' climatology based on that reanalyses' provided vorticity.

We will add the following text:

"RMS daily differences arising from different methods of calculating PV are also small (see Figure A2), no larger than 0.3x10-4 s-1 and mostly better than 0.05x10-4 s-1."

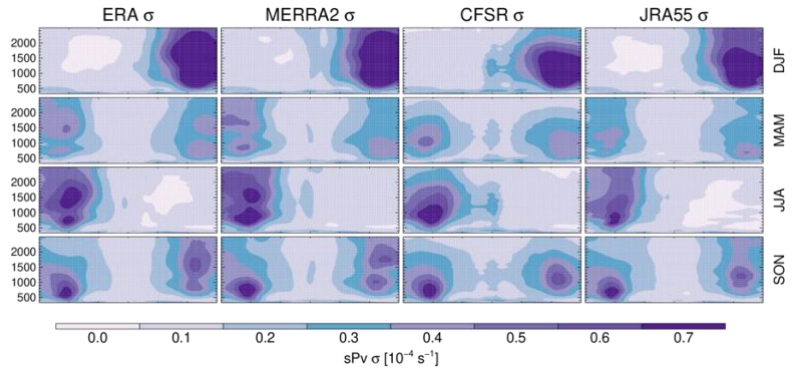
• P6, L8: climatologies "of what" ? Trace gases and aerosols, the sentence will be changed to: to construct trace gas and aerosol climatologies.

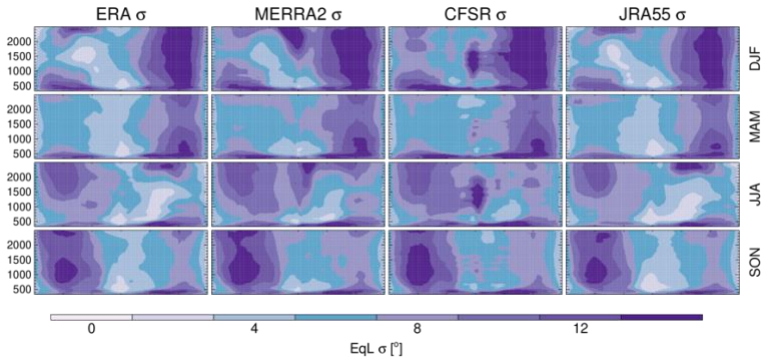
• P6, L13-15: Is there a reference for the EqL computation used here, so that a reader may be able to look up the computation in detail?

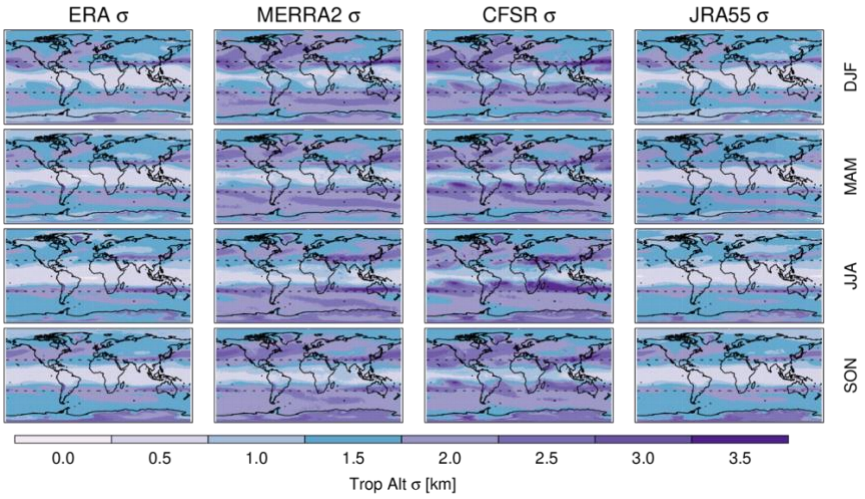
We will add:
"EqL is computed as,
Eql = sin-1 (A/2piR^2 -1)
where A= A(q) is the area in which PV is less than q on a particular isentropic surface, and R is the radius of the Earth.
EqL is computed using the 0.5 gridded PV fields using a piecewise constant method, where the PV value is assumed constant within each grid cell. **Simply, for each PV value, on a given isentropic surface, we sum the areas for all grid cell with smaller field values**. Further, EqL is only …."

• P6, L17: This is potentially the only greater question which I have. Generally, when you speak of variability, here it is talked about the variability along the polar vortex edge, is this a variability caused by the fact that the reanalyses differ in the representation of the atmospheric features among each other, or because there is a large natural variability of the feature. Maybe it could help to also look at the variability of the shown quantities in individual reanalysis data sets to show whether these already have a large variability or not.

The variability in the REM is due to both, mostly due to large natural variability of the feature and differing representations among the reanalysis. Below are the standard deviations for the individual reanalysis for sPV, EqL and the dynamical tropopause. As shown there is large natural variability of these features in the individual reanalysis.







We will add the following:

In the sPV section: "This variability arises from to a combination of large natural variability with the slightly different representations of sPV among the reanalyses."

In the Eql section: "The largest variability is found along the polar vortex edges, as well as at the top of the upper troposphere subtropical jet in all seasons (likely primarily related to EqL becoming a less appropriate coordinate near / below the tropopause, e.g., Manney et al.,2011; Pan et al., 2012); **that is, in regions of large natural variability in EqL."**

In the tropopause section: "Generally, the differences are within 0.1 km over most of the globe, **except in regions of large natural variability.** Around …"

• P7, L7: Do you search the tropopause from top or bottom or asked differently do you refer here to the lowest or highest tropopause in the presence of multiple tropopause as can occur in the vicinity of tropopause folds? We added at the end of that paragraph: "Results shown here are for the primary (i.e., lowest) tropopause."

• P7, L29: I wonder whether the parameterizations of orographic GW really have an effect on the tropopause altitude or whether the effect seen here is rather related to larger, resolved GWs themselves above such orography. As far as I know the standard orographic GW parameterizations rather affect higher altitudes by dumping energy at a specified level somewhere in the middle to upper stratosphere and thus affecting the resolved mean flow at those altitudes but not at the tropopause level. Could this here be also a result of the data assimilation, since these are regions with relatively frequent GW occurrences which might be included in radiosonde data which become assimilated?

The reviewer is correct, we do not have enough information to know if the tropopause differences are due to the parametrization differences or due to resolved orographic gravity waves, or differences in

assimilated data that may include gravity wave information. We will modify the text simply to: "Other ~1 km discrepancies can be found over Greenland and over the Andes mountains **and are likely related to orographic gravity waves that are common in these regions (e.g., Leutbecherand Volkert, 2000; McLandress et al., 2000; Wu, 2004; Doyle et al., 2005; Fritts et al., 2010).**"

References:
McLandress (2000) - 10.1029/2000JD900097
Leutbecher(2000) - 10.1175/1520-0469(2000)057<3090:TPOMWI>2.0.CO;2
Wu (2004) – 10.1029/2004GL019562
Doyle (2005) - 10.1175/JAS3528.1
Frits(2010) - 10.1029/2010JD013891

In the summary section the text changed to: "where mismatches in the location of the sharp decrease in tropopause altitude from the tropics to mid-latitudes are so common as to affect the climatology; **over Greenland and the Andes regions that are affected by orographic gravity waves;** and over Antarctica, where conventional input data are most sparse"

• In the discussion of Fig. 7 starting on P7, l24, I wonder how much the shown differences could be related to the vertical grid spacing of the individual reanalyses? These data sets all differ in their absolute vertical grid spacing as well as the interpolation may be dependent on the actual location of the individual model levels in the tropopause region. Maybe it would be worth adding information about the vertical grid spacing of the reanalysis products in the tropopause region/stratosphere.

A column was added to Table 1 listing the UTLS vertical spacing around 1.2km for MERRA2 and 1.km for ERA-Interim, CFSR/CFSv2 and JRA55.
We also added the following sentence at the end of the paragraph: Part of these differences may be due to the slightly different spacing between model levels (1 to 1.2 km apart at these altitudes) and the actual location of such levels with respect to the tropopause.
We also added in Table 1: "the approximate vertical resolutions of the reanalysis fields for their entire vertical range can be found on Figure 3 of Fujiwara et al. (2017)"

• P9, l4: ...smaller THAN the polar cap Done

• affiliation 2 and 6 are the same Affiliation 6 was deleted.