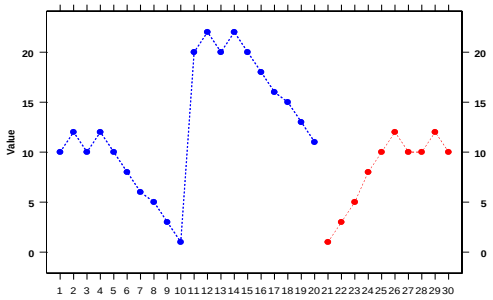
Review of „Occurrence of discontinuities in the ozone concentration data from three reanalyses" by P. Krizan, M. Kozubek, and J. Lastovicka.

The manuscript objective is to search for step changes in the time series of ozone profiles, between 500 hPa and 1 hPa, from three reanalyses, MERRA-2, Era-2, and JRA-55. If the step changes are spurious (i.e. not related to the atmosphere processes) and enough large the trend estimations will be unreliable as forced by changes in technical details of the reanalysis method (e.g. inclusion new satellite data and/or procedure in GCM). Therefore, the subject is important and fits perfectly to the aim of the ACP journal. However, **the manuscript in present form is not ready for publication in the journal and should be rejected**. It requires substantial changes prior any submission.

General Comments.

The authors use the Pettitt test to detect inhomogeneities in the ozone time series. They do not provide reasons for choosing this test and do not discuss its applicability to the ozone time series. There are many other tests to detect series homogeneity (see C. Yozgatligil and C. Yazici, https://doi.org/10.1002/joc.4329). Here, the Pettitt test is used rather mechanically assuming only one change point. The authors are aware that this is not the case for the analysed ozone time series as they discussed possible presence of two change points (~2003 and ~2015, l.369-373). The test works well for cases with a singular discontinuity close to the centre of the series. Below, there is an example illustrating that the Pettitt test fails when multiple change points are present in time series.



Here, the Pettitt test is applied to artificial time series with two change points. 10 values (points in red) are added to the authors' time series (Fig.1 in the manuscript) discussed at the beginning of Section 2 (l.65-101). In this way, a downward jump (at 21th point of the series) to the time series value at 10th point is modelled. P value, which is found at the point with maximum U (12th point according Eq.1), is equal to ~0.15 (according Eqs. 2-3) i.e. above 0.05 limit. This allows to formulate a hypothesis about the lack of discontinuities in extended time series but

the authors' original series (20 blue points) showed clear discontinuity at 11th point with P=0.0092. Thus, the assumption of only one changing point in the analysed time series is crucial for this test performance. Therefore, it seems that the test should be repeatedly performed for the connected parts of the time series, not just once for the whole series.

To make the problem even more difficult, there are possible change points in the ozone time series due to superposition of "natural" ozone forcing factors (QBO, ENSO, Brewer Dobson circulation, the Arctic Oscillations, persistence of lack of sudden stratospheric warming in some periods, etc.). Methods of distinguishing between false and "natural" points of change should at least be discussed in the manuscript.

In the reviewer opinion, an analysis of the change point time is necessary. Just calculation discontinuity occurrences over globe is not enough. History of changes in the reanalyses' methodology has been known and these changes should be linked with the time of step changes disclosed in the time series.

The authors define two types of discontinuity: insignificant and significant. They claimed that only significant ones can erroneously affect the anthropogenic trend values as opposed to the insignificant ones. This suggestion needs justification and should be applied only to spurious discontinuities if they are correctly selected from the ozone time series.

Taking into account all mentioned above problems, the conclusions (especially the last one) are very doubtful.

Specific Comments

It is not clear how the significant differences between in the discontinuity occurrence in the reanalyses are calculated. At first the authors claim (l.193) that "All differences above 1% in absolute value must be regarded as significant because the number of grids is very high (1038240 for ERA5 and 207936 for MERRA-2 ". So, practically all differences shown in Tab.3 are significant. But a few lines later (l.197-199) they state "The variance of DO is at some layers high, so it is the reason why the differences between MERRA-2 and ERA5 are insignificant at the majority of layers". Something is wrong. Please describe the test used to find significance of the differences between DO by different reanalyses. Number of independent cases (i.e. degrees of freedom) is usually used in the calculation of the test significance, not number of all data points, because the observations at neighbouring points are usually highly correlated. A calculation of number of independent data points is not a simple task and depends on the spatial correlation structure of the data.

The reviewer found a problem to understand vertical profiles of DO. I guess (the authors do not provide explanation) that extreme (minimal or maximal) DO in Fig.2 (and in many others Figures) at selected level is shown for the specific month, and average DO is the mean from 12 monthly values. If this is OK why there are so large monthly variations in DO for the fixed layer (it is seen as large distance between max and min profiles, Fig.2). Spurious change step linked with changes in reanalysis methodology should appear simultaneously in all months. Large intra year variability of DO suggests that step changes may include a kind of mixing between "natural" (dynamically driven in dependence of season of the year) and spurious step changes.

The authors define significant step changes in the data using 1-sigma criterion of the difference between the mean values before and after the jump. Here, the reviewer does not discuss if this threshold is enough large to affect the trend calculation. Different problem is how localization of this jump affects trend calculation. It seems that the effect will be strongest when the jump occurs in the middle of the time series. Thus, not only the difference between the means is import here. Presence of multiple step changes affects the mean value after (or before) the jump, so significance of the step change should be calculated taking into account the mean derived from the period between the step changes (e.g. period between 11^{th} and 20^{th} point in the attached Figure). Therefore, the selection of significant step changes needs at least discussion in the manuscript. Searing for a link between spurious step changes and trend calculations requires much more efforts (maybe in new manuscript?) and any statement concerning it should be only hypothesized (and omitted from conclusions) in the present manuscript.

The authors use formula for one-sided probability (Eq. 2, line 87) in the illustration of the Pettitt test. It should be two times larger for two-sided probability, i.e. $P=2\exp(T)$, if the direction of change (up or down) after the step change is not important. Please check if P is used correctly in the rest part of the manuscript.