Review of the paper "Hybrid ensemble 4DVar assimilation of stratospheric ozone using a global shallow water model" by D. R. Allen et al.

General comments:

This paper presents an assessment of the ability of a hybrid 4D-Var/EnKF data assimilation system to extract wind information from the assimilation of stratospheric ozone data, and compares it with that of a standard 4D-Var and a standard EnKF. Understanding if and to what extent ozone observations can constrain the stratospheric wind field in a data assimilation system is of a particular interest considering the scarcity of stratospheric wind measurements.

I find the topic very interesting, and within the profile of this journal. The quality of the study is high, the paper is clearly written and structured, and the results are discussed thoroughly.

I recommend the publication of this manuscript in ACP subject to a few minor revisions and comments (see below).

Specific comments:

1. Page 2 Lines 29-31 (P2L29-31 hereafter): I would have been more comfortable if two different models were used for the TR and DA. As it has been done, the DA experiments are constrained "twice" with the same information (the first time through the background that comes from using in the DA runs the TR model, and the second time by assimilating pseudo-observations derived from the TR itself). I appreciate the statement "results are therefore likely to be overly optimistic", and the care the authors put in designing the experiments, for instance by providing an initial error in the background resulting from imposing the Day=20 T=6h TR fields to be the initial condition for the DA, but I do wonder how significant the differences between the Day=20 T=0 and Day=20 T=6h TR fields really are. This could perhaps be commented in the paper.
2. Related to point 1 above: How responsive is the system? This aspect can help the interpretation of the results on two fronts: *a)* to disentangle the impact of essentially using the same constraint twice (in the sense mentioned above) and possibly quantify the actual wind extraction potential of ozone data assimilation (i.e. understand how overly optimistic the results are); and *b)* to know if at day 10 the system is completely spun up. A comment on this point should perhaps be included as well.
3. Section 2.2 and P5L20: In my understanding, inflation and localization are used in EnKF to address issues (such as filter divergence or long range spurious correlations) that result from under-sampling, i.e. having an ensemble so small in size that is not statistically representative of the state of the system. I'd argue that this does not apply to the case of the "large" ensemble with 1518 members and it would also explain why for that ensemble size the WEP value does not significantly change as function of the localization length (figure 2). The authors may want to comment on this and explain why and in what sense the large ensemble results that used no localization were worse than those with localization (P5L20).
4. P5L4-6: I appreciate the reference to A15, but perhaps the authors can say something more here on the NNMI and make this manuscript more self-contained. Also in A15, instead of NNMI the authors used the acronym NMI. It would be good perhaps to either use a consistent acronym or acknowledge the use of a different name.

5. P7L11-12: I guess the wind vector RMSE is computed as in equation 6 of A15, perhaps a reference can be added.
6. Fig 7: You may want to specify in the right panel caption how the relative difference is computed. I find interesting the reduced efficiency of the 50-member ensemble compared with those with 25 and 100 members. Is the reason for such a result known?