Review of the manuscript "*A stratospheric prognostic ozone for seamless Earth System Models: performance, impacts and future*" by B.M. Monge-Sanz et al. 2021

This study discusses an implementation of a previously described and tested computationally efficient ozone chemistry scheme in the ECMWF operational NWP system. This new scheme replaces the default Cariolle and Déqué formulation. The new BMS scheme models ozone chemical tendencies as sum of linear terms dependent on local ozone concentration, temperature and overhead ozone column anomalies relative to a climatology, with tendency coefficients derived from a full chemistry model. One of the major advantages of the BMS scheme over CD is that heterogeneous chemistry is implicit in the tendency coefficients. The paper demonstrates that the ECMWF system equipped with BMS produces ozone fields that are in better agreement with independent data, including in anomalous polar winter/spring seasons such as during the 2002 Antarctic ozone hole split. Furthermore, when ozone fields are fed back into the radiation code, there are substantial improvements in the representation of stratospheric and even tropospheric dynamics leading to improved forecast skills.

The paper is clearly written, and the results are convincing for the most part. Quantification of impacts of improved representations of stratospheric ozone on forecast meteorology has been a long-standing open problem in our field. I think this work will be of interest to the NWP and S2S communities. I recommend this paper for publication subject to some revisions. I don't have any major criticisms of the analysis, but I do have a number of specific comments. In particular, many of the figures are barely legible, and many of them are not properly captioned. Figure 15 is the most striking example; it contains a large number of acronyms and numbers, almost none of which are explained or discussed anywhere. I also feel that the selection of metrics used (e.g. in Figure 12) needs some stronger justification.

Specific comments

L14. What is this correlation with? ERA-Interim?

L20 "in timescales" → "on timescales"

L45. Why "not yet available"? We have full chemistry models including CCMs. The tools are out there. I think one issue is computational feasibility of having a chemistry model drive ozone in NWP applications and atmospheric reanalyses, although in the latter case there are a few reanalyses that use full chemistry (albeit without radiative feedback, e.g. CAMS, BRAM2)

L73. "previous linear ozone model," → model**s**

L75 expand SLIMCAT (first use)

L85. Is "new BMS" different than BMS discussed above? Why "new"?

L100. “evolves in time according to the following equation”. That’s a little imprecise. This equation contains chemistry only; the full evolution equation will also have transport terms. I suggest rephrasing.

L106. So they implicitly assume a certain level of ODS concentrations, right? Please explain.

L125-126. I don’t think I understand how this type of processes is not missed in Eq. 1. Which of the four terms includes it? The coefficients there are latitude dependent climatological so they don’t know about a polar filament venturing out of the vortex.

Section 2.2. How are the experiments initialized? From assimilated ozone? If so what data are assimilated?

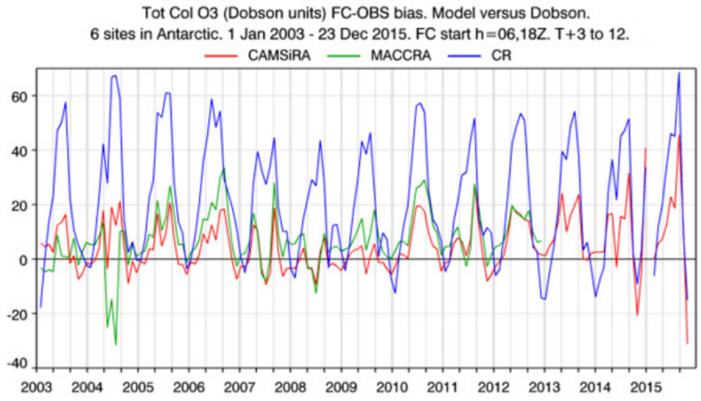
L145. Table 1 only lists experiments at T511 and T159 and only at 91 levels. Other resolutions are for seasonal forecasts listed in Table 2 and discussed in the next subsection.

L168-176. This paragraph lacks any references to published core papers on those data sources, data versions, or on uncertainty estimates.

Sections 3.1 and 3.2 (Figures 1 & 2). Consider also showing and discussing difference panels

L222. “Latitudes over 62°S”. I think “latitudes south of 62°S” would be clearer if that’s what is meant here.

LL224-225. I’m not sure if this statement is supported by Flemming et al. Fig.16, the relevant panel of which I pasted below. Looking at the Flemming et al. figure, in the period 2003-2010 CAMSiRA is biased high by up to 20 DU (not 30) during the summer months, but the bias is quickly reduced during September (onset of the ozone hole) and, actually, becomes negative in October-December. On the other hand, Fig. 3 of this manuscript indicates that in the later months about half of the BMS runs are higher and half are similar to CAMSiRA. I think that the argument would be more convincing if BMS and CD total ozone were compared directly to observations in addition to showing the plots in Fig. 3. Also, what is the average difference between BMS and CAMSiRA in the summer months? Is it comparable to what the Flemming et al. plot shows?



LL235-238. Actually, because all the BMS runs that do close the ozone hole do so within a week or so around 30 November I don't know how much can be said about the variability of ozone hole closure dates.

Fig.5 caption. What are the shaded areas?

LL297-301. I would like to see a more quantitative analysis here. What is the sensitivity of polar winter/spring stratospheric temperature to differences in ozone of the order discussed here (such as between climatological or CD schemes and BMS)? Why should we believe that the cold bias exhibited by the "past versions of the operational ECMWF" model is related to ozone rather that other features of those versions of the model, e.g. gravity wave drag, resolution, representation of subsidence within the polar vortex, etc.? These couple of sentences read as very speculative. I think that either a more substantive analysis is needed (similar to Fig. 7), or these sentences should be deleted.

LL311-321. Some of these improvements are very impressive. Nice!

Fig. 11 the labels on the maps and color bars are unreadable. The font has to be made much larger. It also looks like in several places multiple labels are printed on top of one another.

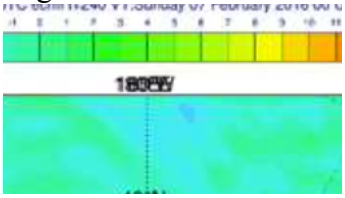


Fig 11 caption. Please state what is getting subtracted from what in panel (d). Also, I think the difference plot would be punchier with a red-blue color scheme.

Fig. 12. Please increase the resolution. The figure is blurry. Also, please add units to the color bar and to the caption.

LL364-371. It's very nice to see this downward signal propagation with increasing lead time but I would like to see some expanded discussion here: I see RMS error improvements 0.01–0.05 (m/s?). How does that compare to typical absolute meridional wind velocities in the regions where these improvements occur? Why were meridional winds chosen as a metric here? Finally, I would like to see a similar discussion of geopotential heights, as this is a more commonly used metric. GPH is shown for the seasonal experiments so I hope it shouldn't be too hard to discuss them here as well.

Fig. 13. Again, it needs a much higher resolution and much larger font. The caption mentions dotted regions indicating significance but there doesn't appear to be any (is that because of poor resolution?).

L376. Are the initial conditions for these experiments from ERA-Interim?

L383. Why SON, not DJF and/or MAM as in the temperature discussion? Is there a rationale behind this particular selection of seasons for different metrics?

L386. Does that constitute an improvement? Would it be possible to verify it against the reanalysis as it was done for temperature?

Fig. 14. The hatching is barely visible. I had to magnify the figure to see it clearly. Also, please consider refining the difference intervals. As it is now, there are large white (i.e. consistent with zero difference) patches marked as statistically significant, which looks strange. Are some of the very small (within +/- 1m/s) differences really significant at 95%?

Fig. 15. I find this figure very hard to read. The caption states "This plot **shows the correlation** between the model experiments and ERA-Interim. Red dots represent the NAO index value from ERA-Interim reanalysis, while blue dots correspond to the ensemble mean of the corresponding model run". This is confusing: either the plot shows the correlation of the NAO indexes, or it shows the indexes themselves. The entire figure looks like it was generated by some standard operational diagnostic software. The ECMWF inner circle may be familiar with the format, but the general reader is not. What are *SNR*, *RPSS*, and *RPSSd*? What does "*CrdOecmfEgi9hS000M001 with 3 ensemble members*" mean? What is "Ratio spread"? What does "Max/Ter/Min" mean? The legends show blue and red dots on top of blue dashed lines, a combination that doesn't correspond to anything in the plots, etc., etc. None of this is explained, or discussed in Section 4.3.1. I really can't tell if this figure shows anything relevant. From the text I learned that the correlation goes from 0.25 up to 0.44 and that's all I got from these plots and discussion. Please replace it with a simple meaningful and properly captioned figure that supports the claims made in this section.

LL395-398. This explanation, while plausible, is not demonstrated. Please, either show that this is indeed the mechanism (preferred) or drop these sentences. It would be nice to see this part expanded and the results demonstrated in detail.

Fig. 16. There are three shaded variability envelopes in these plots and no explanation what variability measure and thresholds are used. Also, please explicitly state that the ERA-Interim line represents the average over the same years if that's the case.

Fig. 17. Please, increase the font size, especially in the contour labels

The experiment names from Tables 1 and 2 are not used very much in the text. Instead, the text uses descriptive names, which is helpful, but it would be good if both were stated. For example, the Figure 1 caption could say "simulated with the new ozone scheme in IFS ("**exp001bms**"; middle) "