Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2020-425-RC1, 2021

Comment on gmd-2020-425

Mathew Evans (Referee)

Referee comment on "A machine learning-guided adaptive algorithm to reduce the computational cost of atmospheric chemistry in Earth System models: application to GEOS-Chem versions 12.0.0 and 12.9.1" by Lu Shen et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-425-RC1, 2021

The computational cost of atmospheric chemistry with atmospheric chemistry transport or an Earth System Models is large and methods to reduce these costs are so useful. The authors have over the years presented a series of papers (Santillana et al., 2010; Santillana et al., 2016; Shen et al., 2020) that attempt to reduce this computational burden by means of (simplifying here) separating the chemistry into fast species for which the differential equations need to be explicitly solved and slow species which can be solved analytically. The complexity of the approach has increased over the years and this paper represents the current incarnation of the methodology.

These technical advances in the numerical methods for atmospheric chemistry transport model (and more generally for geophysical models) are not seen as being sexy science, however, they are essential if these models are to be useful for the wider community. I am thus supportive of this paper and would suggest publication after some clarifications and corrections suggested below.

Major comments.

The figures from the supplementary material should be included in the main text of the paper. For me, it is hard to understand some of the more complex mathematical aspects of the paper without a diagram to support it (e.g. S1 etc). I don't really see why all of the figures from the supplementary material can't be included in the main text. They are not particularly repetitive and I think it would be useful to the reader to see them all.

It would be useful at the end of the introduction to provide a context for what is coming up in the rest of the paper. The lines around 185 provide this commentary about what has been done in the past, the problems associated with those and the method of addressing those which is discussed in the rest of the paper. This would help to contextualize Section 2 which seems like a rather remote set of definitions at the moment. In a few places, the paper feels like it is rather disjointed with one section not necessarily rolling into the next with much cohesion. Perhaps a re-read and a re-think of some of the structure would be beneficial for much of the paper.

The available code is included in a RAR format. This doesn't seem to decompress on my

Mac. Can we have the data in a standard zip format?

Minor Comments.

Line 60. This sounds like the end of an abstract rather than an introduction. At this point, the methodology hasn't been explained or tested so how can it be called chemically coherent, accurate or

Line 87. Can the value of 10 molecules cm-3 s-1 be put into some context? Why was this chosen? Which reactions fall into this category?

Line 99. Giving the number of species (228?) in the reaction mechanism would be useful for contextualizing the 3400 other numbers?

Line 104. "$T_{i,j}$ is the is the number of reactions that include both species *i* and *j." Is that t* as a reactant, products or either?

Figure S1 should appear here to help explain the definition of Di,j. It would be useful to give an explanation for the numbers which are obtained given the chemical mechanism linking Toluene, xylene and Glyxoyl.

The methods used to calculate "distance" appears to refer to the mechanism without any chemistry occurring. In the case of A+B-->C and A+D-->C, the "distance" between A and C doesn't care about the 2 rate constants or the concentration of the reactants. It would appear that chemically minor routes or channels are given the same weight as the dominant routes or channels. Presumably, the ideal way of working out the distance between species would be to explore the mechanism with some chemistry occurring and the weigh the distances between vertices by the flux between species or something like that? This obviously has a significant downside of being much more complex to implements and contextual (the distances would change with the chemical environment). It might be useful to describe this as being an optimal approach, but the approach taken is a simplification of this. Otherwise, it is rather hard to understand why the fluxes have not been used to represent the distance between species?

Line 116. It might be worth including the definition of "long-lived" and fast here.

Line 130. It would be worth pulling in the description of f from the supplementary information.

Line 134. Bring in the figure S2 into the main body.

Line 144. By only considering values greater than 1e6 cm-3 from your statistical analysis you are not excluding many values for typically high concentration species (O3 etc) but you will be excluding a significant fraction of the values for OH shown in Figure S6. It is not appropriate to do this for OH. A significant fraction the grid boxes will have OH concentrations less than 1e6 (the canonical global mean value). I appreciate that there might be increased errors at lower concentrations but if a large number of the global grid boxes have lower concentrations than this, this metrics is being rather selective in the grid boxes that it is using the analyse the model. From Figure S8 it would appear that this excludes evaluation of the model error in large chunks of the surface of the globe for OH?

It's also not clear why the value is set at 1e7 for NO2 in Figure S6? Is this alternative cut off used in Figure S8? Presumably, this is the reason why the polar regions are excluded from Figure S6 for NO2?

Line 158. Can Figure 1 include information about the definition of slow and long-lived for

reader clarity otherwise they are having to flick back to find the definitions used?

Line 182. "different blocks based on similarity of chemical behaviours using a machine learning clustering method." I think this is described in the supplementary material. This should be brought into the main text of the paper or a reference to where this is described included in the text.

Line 185. It would be very useful to have had the contextual information given here much earlier in the paper so the reader can understand what didn't work in the past, what the proposed solution is and how this will be implemented.

Line 209. I'm not sure I understand the comment that iodine reservoirs are inert? Many of them are highly photolabile. This doesn't seem to make sense?

Line 243. When talking about the error this is the RRME? This should probably be clarified. It would also be useful to discuss the implications of the >1e6 cm-3 limit here on the statistics. Is the NO2 value with the 1e7 cm-3 limit as indicated in the supplementary material?

Line 245. Being able to update the chemical mechanism is an important aspect of maintaining the viability of the model in the long term. If there was a significant change to the mechanism the whole process of running the model without the chemical mechanism splitting would need to be done again? The method outlined here is for "on the fly" updates and relies upon the changes being small and the chemical intuition of the person doing the update. It would be useful to explain that the approach described here is for "patching" the mechanism etc. rather than updating the whole mechanism.

It's not clear where new "biogenic VOC" degradation products would go (blocks 7,8 or 9 etc). If an exact mechanism for working out the placement hasn't been found and new species were randomly allocated to 7,8 or 9 (or another mechanism) it would be useful to have that documented.

Line 272. I don't think that it has been demonstrated that it can be ported to different atmospheric models "easily". Relatively minor changes to the chemical mechanism within one model were shown to be able to be incorporated without re-running the whole tuning procedure but I don't think it can be demonstrated that it is easy to move into a different model (CESM etc).

Figure 3. Could be greyed out area showing the number of grid boxes in that category be split into some subcategories: marine boundary layer, continental boundary layer, free troposphere, stratosphere etc to provide some additional information?

Figure 4. This is the performance over what timescale? All of the 1-year timesteps for all grid boxes? It's not clear that the bars indicate the simulation speed up and the symbols represent the accuracy.

Supplementary material. Much if not all of this should be in the main body of the paper.

Figure S4. Is this mislabelled as "anthropogenic blocks", think it should be biogenic blocks? How is the decision made about fast/slow when there are multiple blocks? This could be explained.

Figure S5. I'm not sure that the figure actually shows the "mechanism complexity needed"? It shows the sub-mechanism at each location. I found the % fast labelling slightly confusing as it wasn't that obvious whether the colour scale was describing the Regime or the % fast? Perhaps the % fast could just be removed for simplicity?

Figure S7. How do these curves look if the <1e6 cm-3 restriction in calculating the RRMS is removed?

Figure S8. Can the location where the RRMS has not been calculated due to the 1e6 cm-3 restriction be indicated (grey the area?).

Figure S9. What is the H / L notation indicating?