Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2020-425-RC2, 2021

Comment on gmd-2020-425

Anonymous Referee #2

Referee comment on "A machine-learning-guided adaptive algorithm to reduce the computational cost of integrating kinetics in global atmospheric chemistry 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-RC2, 2021

This paper present some clearly very well thought out and well executed methodologies that are shown to be effective at minimising computational costs for solving a complex chemical mechanism in the GEOS-Chem model. These tool are likely to have substantial benefits for the air quality and Earth System modelling communities. Unfortunately, the writing of the paper is confusing at times, understandable given the complexity of the subject matter but it does distracts from messages it is trying to convey. With a few small changes to the structure and presentation, this will be an excellent paper well suited for publication in GMD.

Major Comments

The biggest issue I find with the paper is the it goes into the nitty gritty of how it partitions different species/reactions into different categories (sections 2.2-2.5) before it explains the overall design philosophy and what these different categories are used for (in section 3.1). The result is that on first reading, I got very confused reading through section 2 and was only able to make sense of it on second reading. This could be improved substatially if either Section 2.1 were expanded upon to give an overview of the whole design philosophy and what each of the the different subcategories (fast, slow species, unimportant reactions etc.) are going to be used for before they are described in detail, or a new Section 2.2. is added giving an overview of all of the developments. Some of this description is in the first paragraph of Section 3.1. Also section 3.1 repeats a lot of the text currently in section 2.1 about the model description and is inapropriate here, therefore section 3.1 should also be editted to avoid repition.

In short, Section 2 should contain all of the descriptions and definitions for the model and the adaptive algorithm for the chemical operator. Section 3 should focus on testing, evaluating and optimising the algorhithm in the 3D model. If these two sections were more clearly defined, the paper would be much easier to follow.

There are also some inconsistancies in language and definitions - I found the use of "slow species" and "slow reactions" (each of which use different threshold definitions) particularly confusing. It would help to be consistent and use "unimportant reactions" for those <10 molecules cm-3 s-1. I have specific comments below to help with this.

In terms of the statistical analysis (Section 2.6) it is important to have a measure of bias as well as error - I would generally be more concerned about a species that has an error of 1% and bias of ~1% (which would imply a consistant error in one direction), than one which has an error of 1% and a bias of ~0% (which would implpy more random error). Looking at Figure S8, the errors in surface Ozone seem to be biased in one direction. That is potentially concerning if the bias is enough to significantly affect tropospheric ozone burden - as ozone is a radiatively important species this could have consequences in an Earth System model. Please also include a normalised measure of bias in Section 2.6.

Finally, there are a number of figures in the supplement which would have been useful in the main paper and I did not understand why they were not in the main paper. I would recomend moving at least figures S1, S2, S5, S6, S7 and S8 into the main paper.

Specific comments

ln 26. Ambigous "it" in: "because it exerts strong forcing and feedbacks...". change to "because chemical and aerosol species extert strong forcings and feedbacks..."

ln 49. Change "guide us build" to "guide us **to** build".

Ln 88. here and elsewhere, please be consistent and use the term "unimportant" instead of "slow" to describe those reactions that are removed with fluxes <10 molecules cm-3 s-1, otherwise it is easily confused with the "slow species".

Section 2.3. If I am to understand this correctly, when species are defined as being "slow" and/or "long-lived", they are not "removed" from the mechanism *per se*, rather they are solved using the analytical approach instead of the 4-th order Rosenbrock solver. Given this, I think you need to make clear here that when you say "coupled system", you mean all of the species and reactions which are solved using the Rosenbrock solver. All of the species that are "removed" or "excluded" from the coupled system, as you say later, are instead solved using the simple analytical approach.

Line 99. Please can you define "distance" qualitively here before the quantitative definition.

Line 112. The term "distance" is overloaded in the text to mean multiple different things. You have defined a new term "Euclidian distance" |Di-Dj|, which is a scaler, rather than

the previous "distance" Di which is a vector. This Euclidian distance is then modified to cluster species together (I would call this something like "clustered distance"). Please make clear that it is this "clustered distance" that is stored in the matrix and used to calculate the cost function Z (eq 4).

Line 113. When applying the 50% factor, is this applied iteratively? i.e. if two species are each others closest pair, will their Euclidian distance be reduced by a factor 0.5*0.5=0.25, as iterating through each species, or is the 50% factor only applied once? Please be clear which approach you are using.

The 50% factor and 5 closest neighbours both seem quite arbitrary, but I think I can see how this approach will cause clustering of species into close families. What was the reason for the use of these two values, and did you test other values?

Line 135. Would benefit from Figure S2 being in the main text here. Looking at Figure S2, I think you can make a fair cost-benefit argument that M=20, N=13 is well optimised as it is on the bottom-left of the 30% contour. It is clear from the contour lines that you get diminishing returns of the fraction of species if you were to increase M and N, hence selecting on the 30% contour seems reasonable. By selecting the bottom-left part of the 30% contour, you are minising both N and M.

Basically, you can be more rigorous in the justification for why you used M=20, N=13 than simply saying you "choose" them.

Line 137-139. Please move this text to the top of section 2.5, as this describes the training set used to derive the submechanisms and blocks.

Line 196. Please again change "slow reactions" to "unimportant reactions". Please clarrify - are the unimportant reactions removed from each of the submechanisms using the original training data, or are they removed on the fly depending on the concentrations of species in each grid cell at each timestep?

Removing the reactions from each of the submechanisms in advance seems like the more efficient approach to me. However, there is a risk that the approach becomes inconsistant if used in different time periods with different chemical conditions to the training data. For example, there will be risk in using submechanisms derived with present day training data in preindustrial conditions. This is relevant for application in Earth System models.

line 229. The full mechanism is by definition not a submechanism. Say that it is the 21st "chemical regime".

Line 236. slow reactions -> unimportant reactions

Figure 1. The line in panel a shows the fraction of species removed from the *coupled mechanism.* Call the slow reactions unimportant reactions.

Figure 3. There are 21 chemical regimes, made up of M=20 submechanisms plus the whole mechanism.

Figure 4. Number of chemical regimes is 21 (M+1), not 20.

Supplementary material

line 12. think Z_2 should be called *f* or *f(M,N)* instead. Use a different term to *D* for the gridcells, because *D* is already used for the distance vectors.

line 18-20. Unclear how this algorithm works, don't know what is meant by "temperature" here.

figure S5. presumably chemical regimes should go up to 21, not 20? Also, its the 21st regime that has 100% of the mechanism, regime 20 has 90% according to Figure 3.

Figure S6. Please include panels showing biases for the key species. If any have consistant/growing biases, that should be commented on (especially concerning for ozone). What value of δ did you use here?