Review for: Intercomparison of Four Stomatal Flux-Based Ozone Damage Schemes in CLM5.0 over China: Implications for Carbon, Water, and Energy Fluxes

This work compares differences in simulated GPP, LE and H output over China when different O3 damage schemes are implemented within the same model. Given the recent increase in O3 damage schemes and greater inclusion of schemes within land surface models, this kind of evaluation is useful and important to model developers.

I am not entirely sure this fits the GMD journal: although the models are compared to observations, the conclusion is that evaluation to observations is not a good metric for assessing skill as the underlying model is tuned to observations and existing model biases obscure analysis on scheme performance. It mostly covers the implications on fluxes and may be better suited in ACP. That said, there are not substantial science conclusions, so I appreciate it is a borderline case.

The paper is quite long given that is it mostly a comparison of outputs. Many sections could benefit from quoting fewer numbers but highlighting the key findings of the analysis instead. At times, the language used is confusing and the meaning is lost.

I do not think that more analysis needs to be done, but some time should be taken to reconsider how your results are presented and to edit to the text. For this reason, I give major revisions although it should be very feasible to address the comments:

Major comments:

There are many moments where a figure is described, but the interpretation is not clearly explained. Specific moments are given below, but each section could benefit from the author considering what the key findings are and how to stop this message getting lost in the text.

Section 3.2 is mostly a list of numbers and abbreviations. It's very hard to follow and understand what points you are trying to make. The table is also quite overwhelming to interpret and could go in the supplementary. I would rethink this entire section by first considering what message you want to convey, then only quoting values and numbers to back up your key messages.

Section 3.3 should have more of an introduction sentence to justify what you are showing and why:

(1) I was confused why you start immediately with sun-shade partitioning of photosynthesis:
- Why do you divide by sunlit and shaded leaves? O3 damage schemes are not calibrated to partition well between these, so the purpose of why you do this needs justifying. For example: What information does it give us that helps understand which scheme is better? What are the impacts of greater damage to shade leaves? Is this something important to measure in future O3 damage studies?
- Why do sunlit and shaded leaves show different responses between schemes? I can only think that it is to do with the coupling of gs-A.
- L279: How does the vertical pattern change? Are you assuming this from the sun-shade fractions? You don't have a resolved canopy, so what does this mean?

(2) The purpose of showing photosynthesis, gs, transpiration and GPP should be explained, and the order of analysis may be improved (e.g. by showing GPP and transpiration first and using photosynthesis and gs to help explain your results):
- Is GPP not just the sum of sun and shade photosynthesis? Are the photosynthesis rates canopy averages or top of canopy values?

- How do transpiration and gs relate – a short sentence motivating why you start by showing gs then move to transpiration would improve the flow.
- What do you learn from comparing photosynthesis then GPP – are patterns the same? Same with gs and transpiration: how does the initial gs analysis help explain the transpiration patterns?

Sect 3.4: This is again a description of numbers and abbreviations and it is impossible to derive the key messages. As with Sect. 3.2 this should be rewritten.

Discussion: The discussion mainly relates to uncertainties, future work and other literature. While this is good, it is missing science discussion on how to interpret your own results. I would be interested to know:

- Do the coupled and decoupled schemes differ substantially?
- Under what circumstances are schemes very different or very similar? E.g. the coupled schemes (Sitch and Ma) seem quite similar in terms of spatial mean changes but divided by PFT show larger differences.
- When might one scheme be better than another? If the schemes produce such different values, how do we interpret results on ozone damage impacts?

Paragraph L355: This is really hard to follow. Are you saying that you see larger GPP decreases than transpiration decreases across all schemes but this is not consistent with all literature? It gets confusing because you compare GPP and transpiration across literature, schemes, coupled and decoupled. I would have separate paragraphs for coupled and decoupled schemes, or for your results and literature values.

Paragraph 371: Here you described the differences in the schemes but you do not say WHY those differences result in different outcomes, so this paragraph does not really give useful information. For example, the relatively uniform PFT parameters of Lombardozzi probably explain why damage is more spatially unform? You write that Lombardozzi13 overestimates damage in forests but its unclear where this came from as you do not have observations of ozone damage. Please reference the text or figure where you derived this conclusion from.

Paragraph L381: You should explicitly address that models are tuned to observations without o3 damage schemes, so adding schemes with large damage factors can decrease model performance without necessarily being incorrect. You should also describe what processes would be affected by integrating O3 damage within the leaf physiological integration and why you think that would help performance.

Minor comments:

<u>Abstract:</u>

L1: first sentence should say: Surface ozone (O3) alters terrestrial carbon–water–energy fluxes… *through vegetation impacts (or similar).*

L23: replace ‘continuous’ with ‘uniform’

L24: how can Lombardozzi13 overestimate when you didn’t evaluate against observations of O3 damage?

L28: You say earlier that spatial impacts are ‘broadly consistent across schemes’ and now that spatial expression depends on choice of scheme. Don’t these sentences disagree with each other?

<u>Introduction:</u>

This is well described and well motivated

Methods:

Table 1 is hard to read: it is hard to see which values belong to which parameter. Can you divide the rows somehow?

Sect. 2.3.1: can you give a table, maybe as supplementary information, on the data ranges of each of these flux sites and perhaps information on measurement heights, canopy height, LAI?

Results:

L164 and other locations: here you use r when r^2 is a more common metric. Later (L217) you use r^2. Can you be consistent?

L170 and other locations: What does it mean to be below the observation arc? Can you describe how to interpret the Taylor diagrams briefly in the first paragraph? Maybe also include if this is hourly data – would daily data be more appropriate for O3 damage modelling?

L176: 'LE retained strong synchrony' – what does this mean?

Paragraph L178: I read this to mean performance decreased because variability decreased, even if seasonal / diurnal cycles follow observations. Can you explain this more clearly?

L189: replace 'milder schemes' with a different word. Also at L250.

L199: I think you can be explicit: Lobardozzi13 degraded model performance and for most o3 damage schemes, biases increase compared to the OFF simulation.

L226: what is the 'spatial ranking of model-observation agreement'? Do you mean that the spatial pattern is the same when different schemes are used?

Figure 6: this could probably go in the supplementary as you quote the key values in the text.

L290: 'darker local cores' what does this mean? Can you replace it with the meaning (e.g. more damage) rather than a description of the figure.

L323: what is 'a more conventional coupling' in this context?

L339: Sentence on 'skill changes do not map directly onto physiological realism' should be explained and justified.

Discussion:

L341: This first sentence: Zhou et al (2026) is an odd way to open the discussion. This paragraph could also be condensed substantially.

Conclusion:

L413: If the comparison to observations is not a good metric to evaluate model skill, why do you include it and what would you suggest instead?

L419: this is very jargony and hard to follow. Many of the claims are not justified with the paper, such as the importance of embedding the o3 damage scheme within the iterative part of the leaf model – if the o3 damage schemes differ so widely I don't see this having a significant effect.

Occasionally, O3 is missing the subscript or written out as 'ozone'.