Introduction and Summary

We thank the two reviewers - Dr. Wickert and an anonymous reviewer (R1) - and two community members - Dr. Armitage and Dr. Madoff - for their thoughtful comments on our paper. We are encouraged that community members found the topic intriguing enough to participate in the review process without being asked. This suggests general interest for the core ideas we want to share with our community.

Despite this interest, it seems our paper in its current form did not clearly communicate our ideas as we intended. We included topics that some reviewers thought were unnecessary while excluding topics that other reviewers thought were necessary. To extremely briefly summarize how we interpret each reviewer's comments:

R1 : Reject the paper for many reasons including that the paper covers too many topics but not the ones that this reviewer thinks are most important. R1 also argues that what we suggest should become community practice is already LEM community practice (as discussed below, we disagree with R1's assessment of what is already LEM community practice). R1 suggests that we present mispractices that are not currently made in the LEM community.

Dr. Wickert : The paper is too long and yet doesn't give enough explanation of some of the surprising results. Similarly to R1, Dr. Wickert argues that some of what we say should be standard practice is already known by our community. However, Dr. Wickert likes the idea of a call to the community to develop benchmarks, just not our current form of it.

Dr. Armitage : There are many issues with LEMs that need to be discussed. If he were to prioritize issues with LEMs for benchmarking, Dr. Armitage might focus on a different set of issues than we highlighted (platform-specific numerical errors, grid resolution effects). However, Dr. Armitage seems to generally like the idea of a call to the community for clear benchmarks and user protocols.

Dr. Madoff : Although Dr. Madoff focuses on an even different set of issues than Dr. Armitage or we focused on, Dr. Madoff also seems to support an open discussion on LEM benchmarking. Dr. Madoff went even further to such other topics beyond benchmarking that the LEM community should address.

We find the fact that all the reviewers, in one way or another, liked the idea of community standards to indicate that this is a topic that needs to be discussed. However, our presentation and example choices did not impress any of the reviewers. As described below, our plan to deal with these revisions will effectively remove most of the components that both the formal reviewers and community commenters took issue with, and such, a formal rebuttal of the main points summarized above is not really warranted. Instead, we describe below our plan for a revised submission.

Moving forward on this submission

We accept that we may have tried to do both too much and not enough with this manuscript. Ultimately we would like to write a paper that will compel our community to agree on benchmarks and best practices for developing and using LEMs. However, in trying to motivate that, we have not presented what any of our reviewers think is a motivating case. In some ways that was part of our point - we (the authors) should not decide how to benchmark but the community should. However, we were not successful at getting our point across.

We would like to revise this submission by cutting out most of the manuscript and presenting only the results on time to steady state. This would allow us to resubmit something that better fits the description of a “short communication.” It would also allow us to highlight these results, which we think are extremely important but were likely overlooked in our original submission. Based on discussions with community members outside of this review process, we think that the time to steady state results will be interesting and useful for many who use LEMs. Focusing on these results would allow us to more fully discuss why these results leave some of our previous assumptions in LEM studies on shaky ground. This would also open a discussion on the scope of what LEMs can and cannot do.

Notably, the time-to-steady-state comparison is a concrete example that motivates the need for the type of LEM benchmarking and intercomparison we hoped to motivate with our initial contribution. By focusing the revised contribution on this portion of our initial paper, we expect we will be able to document one example of why a community effort around benchmarking would be valuable.

Our general sense from R1 is that "mistakes" like we illustrated are not generally made (e.g., use of timesteps that are longer than stable under a courant condition). We disagree with this, but at the same time, we do not want, or feel it would be productive, to write a paper that catalogs the mistakes made by others in published work as motivation. However, to effectively rebut one of the primary criticisms of R1 would essentially require us to do such a cataloging. By writing a short contribution that is more focused on one specific issue, i.e., the variability in the time to steady-state in our experiments, we believe that we can use this as an illustrative example of the types of mistakes that can be made, explain how such mistakes fit in the context of the literature (without criticizing previous work), and use our own "mistake" as motivation for a commentary and call to the community on LEM benchmarking and best use. This new strategy will also allow us to more fully explore the dynamics of the variability in the time to steady-state observed in our experiments. As part of this revision, we would also then remove much of the content that R1 especially found superfluous, i.e., reviewing and definition of terms, etc.

Finally, we would like to thank the editors - Dr. Wolfgang Schwanghart and Dr. Andreas Lang - for helping us navigate the ESURF submission and review process. We recognize their sustained voluntary contributions to ESURF and our community.