Maximilian Gelbrecht
Earth System Modelling Group
TUM School of Engineering and Design
Technical University Munich
Lise-Meitner-Straße 9, 85521 Ottobrunn, Germany
email: maximilian.gelbrecht@tum.de

Munich, December 19, 2022

Dear Editors of Geoscientific Model Development,

Please find a point-by-point response to the comments by both referees below. The comments of the referees are shown in blue font.

Review 1 We thank the Reviewer for the positive assessment of our article. With regards to the inclusion of benefits for Data Assimilation (DA) and NWP, we can fully understand the reviewer's suggestion to optionally include it. In fact, we also discussed this within our author group before. Ultimately, in this article, we want to focus on Earth System Modelling and the benefits and challenges of differentiable programming therein. In the revised version of the article we added a paragraph to the benefits section, in which we also mention the potential and benefits for DA and NWP, similar to the already included references in the adjoint section. However, we would welcome the opportunity to extend this into a possible follow-up article, for which we would need the expertise of the reviewer or other experts like Alan Geer, as DA and NWP are not within our core areas of expertise.

We are aware of the ongoing research of the team at Google that the reviewer mentioned. To our knowledge they haven't published it yet. The involved scientists like Stephan Hoyer and Dimitri Kochkov did however publish noteworthy papers on scientific machine learning and how to integrate knowledge into machine learning methods before. We included their paper on machine learning accelerated fluid dynamics in the revised manuscript.

Review 2

- Many relevant works are not cited. There are several PDE software packages that employ differentiable programming: FEniCS, Firedrake, and devito, to name a few in chronological order. While not originally built to use differentiable programming, it is also possible in deal.II using dual numbers. There are yet more software packages built on these tool kits for modeling individual components of the earth system, for example Gusto, Thetis, icepack, VarGlaS. Granted, no one has built, say, a coupled atmosphere-ocean GCM using these tools, but they're worth mentioning nonetheless The biggest omission regarding differential programming for PDE solvers is Farrell et al (2013), Automated derivation of the adjoint of high-level transient finite element programs. This paper won the SIAM Wilkinson Prize for Numerical Software in 2015.

- We thank the reviewer for this comment. So far in our article we didn't go into detail on the different discretisation techniques that ESMs use. As far as we know, all of the projects that the reviewer lists here are concerning FEM methods. FEM methods are just one possible discretisation technique: (pseudo-)spectral, finite differences, finite volume and other forms of discretisation also all play a role for ESMs. We consider it therefore outside of the scope

of the article to go into a lot of detail on FEM solvers. Even regular ODE solvers can profit from determining the Jacobian more efficiently and precise via AD. However, in our revised article we added a paragraph on discretisation techniques in general. In this paragraph we mention that one can realize differentiable ESMs independent of the chosen discretisation method, as e.g. demonstrated by the Farrell paper that the reviewer suggested, which shows that differentiable programming can also be applied to FEM models. Additionally, in our revised article we also include additional references showcasing the prior research on combining ML techniques with ice-sheet models. We also added further references e.g. to research on combining fluid dynamics with machine learning.

- 32, "Modern AD systems are able to differentiate most typical operations that appear in ESMs": What about flux or slope limiters? Do you believe in discretize then optimize, or the other way around?

- In general, slope limiters can also be part of differentiable ESMs. See e.g. Fikl et al (arXiv:2209.03270v1) for an adjoint based optimization including slope limiters. There are slope limiters that are differentiable in a mathematical sense, and some that are not. AD can differentiate through control flow, so that even slope limiters like minmod that are not differentiable in a mathematical sense might be possible to use. However, this will depend on the practical implementation of the ESM component that includes the slope limiter, its solvers, discretisation and the way the gradient is computed. We are not aware of research investigating this in detail and in the revised manuscript we therefore added comments on slope limiters in the Challenges of Differentiable ESMs section.

- 40, "Third, additional information from observations can be integrated into ESMs with Machine Learning (ML) models." I'd say that ML tools enable you to construct very complex statistical models and train them with the data you have, but ML as such does not somehow enable you to integrate more information from this data into process-based physical models of the earth system than you could with a more old-school statistical system identification or parameter estimation viewpoint. This is classic information theory, see Kullback's 1958 book.

- We thank the reviewer for this comment, but it is not quite clear to us to which old-school statistical system identification or parameter estimation methods they refer. Machine learning methods like artificial neural networks are also not really new. They built upon statistics and optimisation theory like many other methods. ANNs however do provide extremely flexible universal function approximators that, through their very high capacity, are able to model more complex behaviour than many other methods. In our article we also cite various papers, e.g. the work from Um et al., Yuval et al., and Rasp et al., which showcase how ANNs can be used to improve a more traditional subgrid parametrisation. The point we were trying to make here is that, once a process-based ESM (component) is formulated such that it is automatically differentiable, it will also be much easier to seamlessly combine it with ML components; in addition, optimizing both the parameters of the process-based component and the parameters of the ML part will only be possible if both are automatically differentiable.

- 91-94, "Artificial neural networks (ANNs) can be seen as a subset of these models, but differentiable programming goes far beyond these building blocks": A lot of the wording here is conflating what problem you're trying to solve with how you're trying to solve it. Fitting the parameters of a model, whether it's an ANN or process-based physics model, is the answer to the "what" question. There are many ways you could solve this fitting problem. You could use derivative-free optimization methods – it's not a very good idea,

but you could do it. Using gradient-based optimization methods is the answer to a "how" question, and using AD to compute the gradient as opposed to deriving it on pen and paper (which you can still do for some PDE models) is a subset of that "how" question. The fact that you can differentiate through control flow or user-defined types is definitely a compelling reason to use AD. You do address this and quite well in section 4, but it's really important to make the distinction clear.

- We thank the reviewer for their careful review of this section; indeed we should have made this clearer. In the revised manuscript, we revised this section. The first two paragraph are concerning the "how", and the last two paragraphs the "what" from the perspective of an Earth system modeler.

- 170: I think it's worth making a bigger deal out of the fact that you can get the second derivative so easily with AD. It's often painful but still possible to manually derive a first-order adjoint model, but going to second order by hand is really atrocious.

- We agree with the reviewer that computing second derivatives can have considerable benefits in theory, and we do mention this in a number of places in the manuscript, e.g. in the overview figure. In theory, this can indeed have considerable benefits. However, this also comes at a cost. In particular, computing the Hessian can consume too much memory to be worth considering. Often, methods rather try to estimate a Hessian-vector-product in ways that don't actually need second derivatives at all. That being said, if more models and tools are able to compute second derivatives easily, it is possible that more algorithms will be developed that might avoid the huge memory cost of the full Hessian and actually use proper second derivatives. Therefore we added a comment on Hessians to the manual vs automatic adjoint section.

- 175: Here it's worth citing some of Noemi Petra's work, including here paper on stochastic Newton MCMC as well as her more recent work on hIPPYlib.

  We thank the reviewer for pointing us to this work and we added it to the revised manuscript.

On behalf of the authors,

sincerely,

Maximilian Gelbrecht