Review response for Universal Differential Equations for glacier ice flow modelling

The Authors

September 20, 2023

Dear Reviewer ,

We thank you for your useful comments and suggestions. Here we have elaborated on some of your points. Original *reviewer response is in italic*, while the authors response can be found in blue.

The manuscript presents an approach that combines inverse methods ("data assimilation") with machine learning-based parameter estimation using the concept of Universal Differential Equations (UDEs) or Neural Ordinary Differential Equations (N-ODEs). The approach has the potential to enable a wider and more rigorous use of diverse observations to constrain or calibrate physical models. For it to work, the notion of differentiable programming becomes key, i.e., the PDE representing the physical model and the UDE embedded within the PDE need to be "differentiable" in the sense that an adjoint model needs to be generated. This adjoint model serves to compute the gradient of the "cost" or "loss" function with respect to uncertain/unkown parameters. The gradient, in turn, is an essential ingredient for gradient-based optimization. A versatile approach to generate the adjoint of the PDE/UDE system is by means of automatic differentiation. The approach is demonstrated by calibrating the glacier flow model ODINN.jl on a range of glacier observations from the Randolph Glacier Inventory. The example is simple, meant as a prototype demonstration of the method. The PDE consists of the SIA model for glacier flow. The parameterization within the SIA to be improved upon is Glen's creep parameter A and its functional representation on the climate temperature normal T. Parameter(ization) A is replaced by a neutal network (NN), and the task is to estimate the NN parameters θ. The manuscript concludes with a discussion of challenges and opportunities.

The manuscript presents an exciting approach of combining physics-based inverse modeling with machine learning-based approaches to improve the way that observational data may be used to rigorously constrain or calibrate models. The seamless integration of PDEs and UDEs for model calibration or parameter learning represents in my view a promising applications of machine learning for surrogate modeling in combination with physics-based models. I recommend publication after minor revisions. My comments are largely intended to improve the manuscript's clarity.

Main (but minor) comments:

- *line 136-137 and eqn. (3) "Instead of considering that the diffusivity $D(\theta)$ is the output of a universal approximator, we are going to replace the creep parameter A in Equation (3)..." I agree with this approach, but perhaps a sentence to motivate this choice might be warranted.*

  We explained this in the last paragraph of the section with the following sentence: "Nonetheless, this simple example serves to illustrate the modelling framework based on UDEs for glacier ice flow modelling, while acting as a platform to present both the technical challenges and adaptations performed in the process, and the future perspectives for applications at larger scales with additional data.".

- *Also, eqn. (7) seems to be the crux in connecting the PDE with the UDE. It might be worth, then, to be very explicit that $A(\theta)$ becomes the surrogate model embedded within the PDE. To connect eqn. (7) with the general eqn. (5), perhaps relate $A(\theta)$ with the universal approximator $U(\theta)$ defined in eqn. (5).*

  Thank you for the suggestion, we have added an observation with this point.

- *line 141: "...by prescribing an artificial law for which a reference dataset is generated..." This is perhaps somewhat imprecise. As stated, it gives the impression that a reference dataset is generated for A, when in fact, the actual data set to be used (and generated via A) are the velocities obtained as solution and used in the loss function.*

  We agree this is rather confusing. We have rephared this sentence and added it in the following format: "Given this artificial law for $A(T_s)$, a reference dataset of velocities and ice thicknesses is generated by solving the SIA equation (2)".

- *lines 142/142: "...we have used the relationship between ice temperature and A from Cuffey and Paterson (2010), and replaced ice temperatures with a relationship between A and T" This is a crucial methodological step in this work to generate the reference data. This step should be described more clearly. I think what is done here is to provide a synthetic solution for a known functional form A(T) described in Cuffey and Patterson (2010), which then will be attempted to be inferred. I feel this relationship should be reproduced here to make this paper more accessible and self-contained.*

  We thank the reviewer for the suggestion of making the synthetic relationship between A and T more explicit in the manuscript. However, we believe the actual relationship between these variables is not really important to illustrate the behaviour of our model and could mislead the reader on the specifies of this relationship. Furthermore, it is important here to remark that we are using proxies for the ice temperature and not the ice temperature directly, making the decision of the fake law even more arbitrary for actual scientific purposes.

- *Figure 2: Perhaps two such schematics should be drawn, one using the functional form of A as taken from Cuffey and Patterson (2010) to generate the reference data (with added noise to A), and the other one as is(?)*

  As we discussed in the following point, we believe it is much simpler to keep the current figure, which easily encapsulates both concepts, and gives a more general overview of the modelling philosophy.

- *The box for the "Physical law": wouldn't it make sense to state A as a function of the NN, instead of D, because the rest of the functional form of D remains unchanged?*

  Even if the part of the diffusivity D that is being modelleled using a neural network is just A, we believe this formulation is more general and communicates better the idea of the model. This scheme is more general and can accommodate more variations of the presented parameterization.

- *Also, why is there an arrow from the PDE solver to the Mass balance box?*

  The reason for this is that mass balance models depends of the ice surface elevation: higher altitudes are associated to colder climates with more accumulation. Then, the different values of the diffusivity D_θ will lead to different solutions of ice thickness H that then will have to be fed again in the mass balance model to recalculate the accumulation/ablation rate. This step is essential to correctly take into account the topographical mass balance feedback explained in Zekollari et al. (2020) and Bolibar et al. (2022).

- *line 223 and again 523: "checkpointing": Accurate checkointing is not actually an interpolation between stored checkpoint. AD-related checkpointing methods are designed to balance storing vs recomputation with the aim to recover the required state exactly. Griewank and Walther (2008), chapter 12 cover this. In the context of the latest implementation in Julia, see Schanen et al. (2023).*

  We thank you for the references and observations for this section, we have added them to the manuscript. We have changed the phrasing of these two sentences based on this observation.

- *Equation (8): A typo denominator, should be $\nabla_\theta A$ ("A" missing)*

  Noted.

- *line 350: "Julia source code is automatically differentiable, ...": This is certainly too strong of a statement. While sometimes claimed by members of the core Julia community, it is currently an ambition but which clearly is not (yet) realized in its generality. To move toward this goal requires further developments in general-purpose automatic differentiation. Arguably, the latest developments of Enzyme.jl (Moses et al. 2022) are a promising step in this direction.*

  We completely agree with this point and we agree that the sentence is rather general and not accurate. We have removed this sentence from the manuscript but leaving the discussion on the different AD packages.

Other minor issues:

- *line 59-60 and again line 67: The statement "Nonetheless, all efforts so far have been applied to the inversion of scalar parameters, i.e. parameters that are stationary for a single inversion given a dataset." is not quite accurate (or I misunderstood it, in which case it may need clarification). The study by Goldberg et al. (2015; see references) inverts for time-varying and spatially distributed open boundary conditions. You may mean something else though with this sentence(?)*

  This is a good point, also raised by Reviewer 1. Following these two comments, we have rephrased the whole paragraph in the following way:

  "These inverse modelling frameworks enable the minimization of a loss function by finding the optimal values of parameters via their gradients. Such gradients can be found by either manually coding the adjoint or by using AD. Nonetheless, all efforts so far have been applied to the inversion of scalar parameters and sometimes their distributions, i.e. parameters that are stationary for a single inversion given a dataset. This means that the potential of learning the underlying physical processes is reduced to the current structure of the mechanistic model. No changes are made to the equations themselves, with the main role of the inversions being the fitting of one or more parameters already present in the equations. To advance beyond scalar parameter inversions, more complex inversions are required, shifting towards functional inversions. Functional inversions enable the capture of relationships between a parameter of interest and other proxy variables, resulting in a function that can serve as a law or parametrization. These learnt functions can then be added in the currently existing equation, thus expanding the underlying model with new knowledge."

- *line 118: The use of the term "observed" may be misleading in the present context. While true in general, the "observed" data used in study are actually simulated reference data (from a known functional relationship). It may be worth clarifying this. Also, to be notationally clear, these "observed"/simulated reference velocities are the elements*

V_1^k in eqn. (6), right?

We have changed "observed" to "target" to be more precise about this point. Regarding the second question: that is correct, the velocities described as observed/target are the V_1^k (denoted as u_1^k in the new version of the manuscript) in Equation 6.

- *line 128: In the case where the ω_k are the diagonal elements of a covariance matrix, this would essentially amount to a simple weighted least-squares optimization problem (with weights equal to the the inverse variances).*

  Indeed, the weighted least square problem can be interpreted as a maximum likelihood model where we include the information about the covariance matrix. However, notice that here the optimization is still carried with respect to the weights of the neural network. However, here the weights $\omega_k = \|u_0^k\|_F$ are not playing the role of variances. Instead, they control the differences in observed velocities due to changes in A, so the ideal desired weights should account for variances on A instead of the parameter models. This is an interesting point to explore in future work.

- *line 179: "... output of the prescribed law ...": Just to clarify: the noise is added to the output of the prescribed law $A(T)$, i.e., one generates $A(t) + \epsilon$, with ϵ being the added noise?*

  This is correct.

- *Next sentence: "This setup is used to compute the reference solutions, ...": It might again be worth clarifying that these are the reference solutions on the velocities V_1^k used in eqn. (6).*

  We have changed this sentence to "reference synthetic solutions" to emphasize this point. We have also added the reference to the u_1^k in Equation 6 to make the link more clear.

- *Section 3.3 It might be useful to replace everywhere "mass balance" (and "MB") by "surface mass balance" ("SMB").*

  Thank you for the suggestion. We have added "surface" in the manuscript when referring to the surface mass balance.

- *line 197: "H matrix": not defined what this refers to.*

  The ice thickness H has been introduced in section 2.1 in order to write the SIA equation. We have added "glacier ice thickness H" to this sentence for clarification.

- Section 3.4, 3.5: I suggest merging these two sections under the title "Sensitivity methods, differentiation, and optimization" and convert current section 3.5 to 3.4.3.

  Thanks for the suggestion. We believe it might be clearer to keep these separated. First, to avoid too much depth in the subsections, and second because the optimization methods are independent from the sensitivity methods.

- *line 255: "unrepresented" -¿ "under-represented"*

  Noted.

- *line 257:"2i2c": please add a reference*

  We have added a reference for JupyterHub in this section. Unfortunately, there is no direct way or acknowledging 2i2c, so we removed this mention in the manuscript and keeping the reference to 2i2c in the acknowledgement section.

- *line 263: "with respect TO a predictor"*

  Noted.

- *line 270/271: "until it finds an optimal non-overfitted solution". Could you briefly elaborate how you determined that there is no over-fitting.*

  This is a very empirical observation based on the converge plot shown in Figure 3. Here we can observed how the loss function gets smaller but at the same time that the obtained law for $A_\theta(T)$ is smooth. Over-fitting the solution in this context will mean that the NN will over-fit the values of $A(T) + \epsilon$, that is, the values of A with the added noise. The fact that the model does not try to fit the exact values of A is a good indication that the learned curved is robust to noise and it is estimating the actual value of $A(T)$.

- *line 299-301: How should one interpret this? Shouldn't the functional form $A(T)$ be different in both cases then? I.e., the error in the SMB, should be compensated by the NN model of $A(T)$?*

  This is what we mentioned in the discussion. This is mainly due to the robustness of ice surface velocities to changes in surface mass balance. If one used H, the ice thickness, then the signal would change too much for the network to be able to recover the function. This robustness is linked to the observations, not so much to the method (i.e. the NN) itself.

- *line 306-308: "This weak dependence...": This is certainly a valid perspective. However, if one generalizes the problem to also use altimetric data to constrain ice height, then* *this no longer applies and weak dependence on SMB may become problematic(?)*

Indeed, this is a good point. As we mentioned in the point above, this is good news for rheology/basal conditions inversions IF ice velociy data is available. Altimetry data could also be used in conjunction (we haven't investigated this, though), but it would then be affected by the SMB signal.

- *Subsubsection 5.2.1: Automatic differentiation is independent of SciML. I would give it its own subsection.*

  The main idea behind this division is to divide the discussion in the two main topics of our research: (1) on the one hand, we have glaciology, and the domain-specific science related to the physical laws and the solving of the differential equations. (2) On the other hand we have the Scientific machine learning topics, which in our view also encompass AD methods.

- *line 355: There is a newer reference available: Moses et al. 2021*

  Noted! Thank you for pointing this out.

- *line 373, 380 (and elsewhere): What you call "backward mode" is usually referred to in the AD community as "reverse mode" (see, e.g., Griewank and Walther 2008).*

  This is correct. Although the two words refer to the same behaviour, we have changed "backward" for "reverse" to stay with the already existing convention. We keep the cases where the use of backwards is semantically more correct, for example "These can be classified depending if they run backwards or forward with respect to the solver."

- *line 389: "the use of AD for optimization..." Not quite right. What you probably mean is " the use of adjoints for optimization". AD is merely a way (albeit a powerful one) to generate code, i.e., the adjoint operator, that efficiently computes the required gradient.*

  This is a good point. We have repharased this to "the use of adjoints for sensitivity analysis integrated with AD tools for optimization and the properties of the landscape generated when using numerical solvers has not" to include the use of adjoints but also emphatized that these also need AD to effectively calculate the adjoints.

- *line 399-401: Sentences: "The gradient calculated by making variation of the parameter θ capture the variations of $Err(\theta, hyper(\theta))$, which lead to spurious gradients. On the other side, automatic differentiation compute gradients using one single evaluation of θ, meaning that it differentiates only the term $Solver(u0, t0, t1, \theta)$, being robust to the error term." I would love to see this formulated more clearly. As formulated, I find it not easy to understand.*

We thank the reviewer for rising this point. We have re-write this section to make it more clear and understandable.

- *line 402: obtain − > obtained*

  Noted.

- *Figure A1: Should also define what the points indicated as diamonds are. Also, there exist a classification of numerical discretizations in the atmospheric modeling community, going back to Arakawa and Lamb (1977) Does your grid correspond to any of these?*

  This is a very interesting point. The staggered grid used here corresponds to Scheme E in Arakawa grids, with the difference than here the quantities we evaluate are different (instead of velocities, we evaluate surface gradients; instead of depth, here we evaluate diffusivity). We thank the reviewer for making this interesting connection.

REFERENCES:

Arakawa, A and Lamb, V.R., 1977: Computational design of the basic dynamical processes of the UCLA general circulation model. Methods in Computational Physics: Advances in Research and Applications, 17, 173–265. doi:10.1016/B978-0-12-460817-7.50009-4

Goldberg, D. N., Heimbach, P., Joughin, I. & Smith, B. (2015). Committed retreat of Smith, Pope, and Kohler Glaciers over the next 30 years inferred by transient model calibration. The Cryosphere, 9(6), 2429–2446. https://doi.org/10.5194/tc-9-2429-2015

Griewank, A. and A. Walther, 2008: Evaluating Derivatives. https://epubs.siam.org/doi/book/10.1

Moses, W. S., Churavy, V., Paehler, L., Hückelheim, J., Narayanan, S. H. K., Schanen, M. & Doerfert, J. (2021). Reverse-Mode Automatic Differentiation and Optimization of GPU Kernels via Enzyme. SC21: International Conference for High Performance Computing, Networking, Storage and Analysis, 00, 1–18. https://doi.org/10.1145/3458817.3476165

Schanen, M., Narayanan, S. H. K., Williamson, S., Churavy, V., Moses, W. S. & Paehler, L. (2023). Transparent Checkpointing for Automatic Differentiation of Program Loops Through Expression Transformations. Lecture Notes in Computer Science, 483–497. https://doi.org/10.1007/978-3-031-36024-4$_3$7