Review of: "Deep learning applied to glacier evolution modelling" by Bolibar and co-authors

1 General comments

In this study, Bolibar and co-authors train various machine learning algorithms to compute the surface mass balance (SMB) of 32 glaciers in the French Alps. The paper is well written, timely (machine learning is a trendy topic and will continue to be so in the future), and is an interesting read. Its focus is solely on model development and performance, without application or discussion of model output: therefore, "Geoscientific Model Development" could have been a better venue for such a study.

I am sympathetic towards the main premise of the paper, which is to demonstrate the suitability of deep-learning for SMB modelling, and the open-source tools provided with the paper further increase the relevance of the paper as an example for future studies to build upon. However, I have some concerns with the current (unclear) focus of the study and certain methodological aspects, which I believe need to be addressed before publication.

1.1 GC1: the use of glaciological predictors

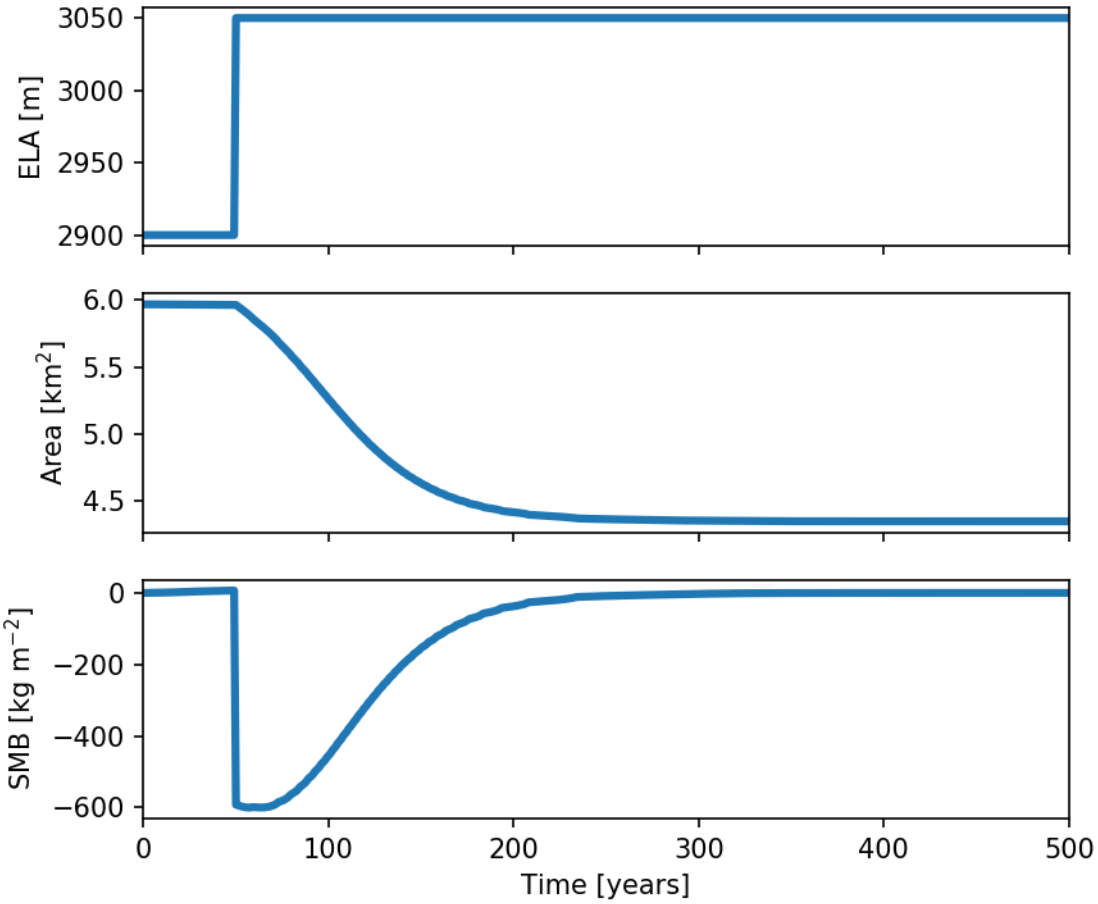
Currently, the statistical models have the possibility to train on certain topographical predictors (more specifically: Mean glacier altitude, Slope of the lowermost 20% glacier altitudinal range, Glacier surface area). These predictors are time-dependant and extracted from DEMs and inventories at various times during the study period. Regardless of the fact that these data are very unlikely to be available in such a precision for many other regions of the world (and certainly not for past and future glacier states outside the observation period), using them as explanatory variables poses a serious conceptual problem: these variables are meant to be **simulated** by the full model (SMB + glacier evolution), thus contradicting the need for a glacier evolution model in the first place. I see three ways out of this chicken and egg problem, all with drawbacks and likely to affect the accuracy of the model:

- use time-independent predictors such as a constant area (probably a bad idea because this will raise model validity problems for longer simulations)
- show that your full model is able to simulate those, and then use the modelled ones as input data for the next year in the "model application period" (i.e. your statistical model will have to be called in yearly time steps). This is possible but will require some thinking about how to validate the procedure.
- don't use them at all (simplest)

Regardless of your choice, the study will have to be adapted to this change. Note that I saw that the predictors weren't chosen by the Lasso model, but: (i) you don't know if they aren't chosen by the cross-validation models, and (ii) because I'm unsure how the predictor selection for the ANN really works I don't know if they play a role there.

1.2 GC2: glacier wide mass-balance

The model is trained to reproduce glacier wide mass-balance (or "specific MB"). Glacier wide mass balances are dependent on the altitude-area distribution of the glacier and therefore are not only dependant on climate but also on the glacier's dynamical response to current and past climates. This has been discussed elsewhere and in another context (e.g. `https://doi.org/10.5194/tcd-4-2475-2010`) and there are good arguments for both sides, but I still can't believe that predicting glacier wide mass-balance is a good idea for a glacier evolution model. For example, consider this idealized glacier response to a step climate change:



Source and context: `https://oggm.org/2017/10/01/specmb-ela`

In this perfectly linear SMB framework (linear gradient, linear response to step change), the simplest of the statistical models could simulate the fixed-geometry SMB (or even any point SMB) *perfectly*, but it would completely fail to simulate glacier-wide SMB, which requires knowledge about past glacier states and evolution. This is an extreme case, but still raises questions about this study (even in the relatively short period considered here).

The large-scale glacier evolution models I am aware of use either an altitude-dependant SMB (e.g. OGGM, GloGEM, PyGEM) or parametrize this non-linear response in their SMB model (Marzeion et al., 2012). I think that it is too late to change this in your framework at this time, but I strongly recommend to explore other approaches for future studies based on ALPGM. If your model ever intends to simulate many glaciers over long periods, I think that this effect should be treated explicitly (or it should be shown that the "black-box" ANN can properly deal

with the full-glacier problem as suggested in the discussion section). Regardless of your choice, this point needs to be discussed in the paper.

1.3 GC3: focus of the study

In my understanding, this study attempts to make three points:

1. Machine learning (and deep-learning in particular) is a useful tool for glaciology
2. Introduce and validate a new SMB model based on deep-learning
3. Introduce a new glacier evolution model (ALPGM)

While I think that the study is fairly successful for points 1 and 2, it does not succeed for point 3. My concerns about point 3 are strongly driven by methodological considerations (GC1 and GC2 above, the use of a perfectly fitted Δh method impossible to validate, and the lack of proper out-of-sample validation of the full ALPGM model). This confusion about the goals of the study also make the paper's introduction and title quite confusing. I would much rather see this study focus on point 1 and 2 (for which you provide tangible results and arguments) and remove point 3 (and the corresponding section "3.3 Glacier geometry evolution: validation" which, in the authors own words, isn't the main focus of the study). Removing point 3 would help to focus on the strength of the current version of ALPGM as a mass-balance model. If the author's choose to keep point 3, then I have several concerns about whether ALPGM really is a glacier evolution model (yet).

2 Specific comments

Abstract L22 : "for past and future climates.": remove "future", since this has not yet be demonstrated.

P2 L6-16 : although it is tempting to classify the models like this, I think that this list (and several other parts of the introduction) needs more precise definitions and a clearer positioning of the ALPGM model. In this list, you need to differentiate between the treatment of ice flow / glacier evolution by these models (on which your classification seems to be based, but not explicitly so) from the treatment of surface mass-balance (SMB), which is what your study is actually about. The "Physics-based models" that you list in fact often have no SMB module, and rely on external SMB as an external boundary condition in real-world applications. For the sake of clarity and given the scope of your study, I would rather focus on the hierarchy of SMB models (with SEB or even coupled Atmo-SEB models being the more advanced, and temperature index models the simpler models). Please rethink this part of the introduction, as well as the following paragraph.

P2 L34 : "Compared to other fields in geosciences": which ones?

P2 L34 : "the glaciological community has remained quite oblivious to these advances": this is a subjective statement, you need to mention that this is your opinion.

P3 L8 : "but all of them were linear, which are not necessarily the most suitable for modelling the nonlinear climate system": you have a very "statistical" view of linearity here. The

statistical models used in Maussion et al. are linear, yes, but they target individual SEB fluxes which are then transformed to be physical (e.g. by preventing negative precipitation or a non-closed SEB budget) and then used to compute the SMB. As a result, the full model M (as in $SMB = M(y)$ with y the predictors and SMB the target variable) is non-linear. This is important also in the context of traditional temperature index or degree day models, which can be compared to linear models applied to transformed predictors and as such, are also non-linear (e.g. by preventing melt for negative temperature or by transforming precipitation to solid precipitation). This is an important feature: without this non-linearity, they wouldn't work at all.

P4 L15 : "When most glacier models tend to incorporate more and more physical processes (Maussion et al., 2019; Zekollari et al., 2019), ALPGM takes an alternative approach based on data science." Are you talking about SMB or ice dynamics? Your "data-science" is applied to the SMB problem here, and I believe it would be more appropriate to cite models of SEB/SMB in this sentence (e.g. Hock et al, Mölg et al, CROCUS, or similar).

P5 L11 : "leave-one-glacier-out (LOGO) or a leave-one-year-out (LOYO) cross-validation". Congratulations for coining these - I wish we had invented these acronyms earlier.

P5 L29-34 "Although the features used as input (...) will likely have different biases." This paragraph seems out of context here and should be moved to the discussion

P6 L25 : "StatsModel" is spelled "StatsModels"

P8 L1 : "The generated coefficients from the model serve to determine the significant predictors to be kept for the artificial neural network training." is Lasso part of the feature selection process of the ANN then? This raises interesting (and hard) questions concerning cross-validation and the model's real independence from training data. Furthermore, it gives an advantage to ANN over the linear models since their predictors are pre-filtered (see e.g. the "double Lasso" method which makes this an advantage as well). Please comment.

P9 L5-17 : hyperparameters. As a non specialist of "deep-learning", I need to ask: shouldn't this hyperparameter selection also be cross-validated? In Lasso, for example, the regularization parameter could be called an "hyper-parameter" and its selection takes place within the model tuning step, effectively making any external cross-validation (realized by LOGO and LOYO in your case) a "true" out-of-sample validation. What about the ANN hyperparameters? Please comment.

Glacier geometry update You call the geometry update a "parametrization" but in my opinion it isn't: you use an empirical Δh function perfectly known for each glacier since it is individually fitted. A true "parameterization" (like the one used in Huss and Hock 2015) would have the goal to work for any unseen glacier. Currently your model cannot be applied)(or validated) against unseen glaciers.

Figure 4 : Since you have DEMs (and geodetic MBs) from all blue glaciers in Figure 4, can you apply your model to them as well and compare? This would be a good (but partial) out-of-sample validation ("partial" because you still need knowledge about the glacier's Δh).

Glacier ice thickness : To avoid confusion: if still applicable after revision, mention here that ice-thickness are only used for the 2003-2016 test run, and not for the rest of the model workflow.

Glacier topographical variables : from an email question to the authors I know that the topographical predictors (e.g. area, slope) are time-dependant and obtained from various DEM snapshots. This needs to be explained here. Regardless of this missing explanation, this raises questions about the overall applicability of the method to unseen situations (see general comment).

P16 L2 : "For the training of the ANN, no combination of topo-climatic features is done as previously mentioned". I have a hard time finding where this is explained. Is this the part with Lasso? In any case, the predictor selection for ANN needs to be explained here for consistency and to help the reader.

P16 L9 : "Latitude and longitude seem to play an important role when combined with snowfall.". I don't really understand the climatological explanation that follows this statement. If the reanalysis data is accurate, then these east-west and north-south differences should already be in the training data. If anything, these combination of predictors play the role of bias correction - or are the result of luck (which is often the case with many co-linear predictors).

Lon/Lat Predictors : this is a subjective opinion, but I suggest to remove Lon/Lat predictors from the set. They should not explain anything which isn't in the climate and topographical predictors already, and using lon/lat seriously hinders the applicability of the model to larger areas.

Figs 6 and 7 : although visually appealing, the use of different colorscales for the ANN and linear models is misleading. All four plots are exact same and should have the same colorscale, min-max range, x-y axis, etc.

Figs 6 and 7 and corresponding discussion about explained variance : a possible improvement to describe the models errors is to plot binned model error (residuals) as a function of the target variable (here, SMB), or use a Q-Q Plot. It would display in a more quantitative way the non-normal distribution of model errors (visible on the scatter plots by a flattening on both ends of the scatter), further making your point that ANN is better (but not perfect) at reproducing the true variance of the data.

Figure 10 : a striking feature of figure 10 and not discussed in the manuscript is the clear tendency of LASSO to overestimate MB in the second half of the period and underestimate MB in the first half. I guess it is a result of more frequent negative MBs in the second half, which are underestimated by the model with an obvious lower variance, but is this the only reason?

Glacier geometry evolution validation : these results are not too surprising. Since your evolution model knows exactly where mass is going to be removed (based on data going up to 2011). This test is basically a bias test: if the model has no bias, ice is going to be removed at the right place (because you know where to remove it) and the area will be correct, provided that the ice thicknesses are more or less accurate.

P24 L1 : "Even for a 12-year period, the initial ice thickness remains the largest uncertainty": this statement is not supported by your results, since this is the only uncertainty you consider in Fig. 11. Here, you could add model uncertainty by using out-of-sample validation (by training the model with data only before 2003 and using LOGO), or use uncertainty measures derived by cross-validation. The issue with the Δh method raised above would remain, though.

P25 L26 : "we trained an ANN only with monthly average temperature and snowfall, without any topographical predictors". These experiments should become the central component of your study, not the other way around (see general comment).