Fabien Maussion

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, Geoscientic 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.

We would like to thank Fabien Maussion for his thorough, constructive and generally positive review, which raised very interesting questions regarding glacier modelling. We believe that it has helped to revise certain methodological aspects as well as to bring interesting new elements in the discussion. All comments from the review have been responded individually.

Regarding the journal choice, as it will be later discussed in GC3, “The Cryosphere” has been chosen over “Geoscientific Model Development” in order to focus more on the application and feasibility of a machine learning approach for SMB reconstruction and simulation, coupled with a glacier geometry update component, rather than the presentation of a new glacier evolution model. ALPGM has been developed as a tool during the first year of the main author’s (Jordi Bolibar) PhD thesis. Therefore, we believe that the interest of the chosen approach (deep learning SMB modelling) prevails over the general interest of the model and code itself. Since the machine learning SMB modelling approach requires to re-train the model(s) for other regions, the most interesting part is the methodology itself, and not the fitted parameters that are, after all, specific to each geographical region. ALPGM being open-source, can be easily re-used, but a limited amount of time was dedicated to creating easily re-usable code and interfaces, needed for a readily shareable glacier model. This kind of purely software engineering tasks are extremely time consuming, as the reviewer will know from his own experience with OGGM. Therefore, we preferred to discuss the approach itself (the main topic of the article) which can be re-used by everyone (with or without our code) rather than making the presentation a model which is far from being “plug-and-play” applied to other regions. In that sense, the aim of this paper is to present the approach and methodology that will be used in future studies to simulate SMBs and the evolution of glaciers in the French Alps (the goal of Jordi Bolibar’s PhD thesis).

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:

(a) use time-independent predictors such as a constant area (probably a bad idea because this will raise model validity problems for longer simulations)

(b) 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.

(c) 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.

These aspects of the functioning of the model and its specific behaviour appear not to be clear enough in the manuscript, so we reworded parts of the manuscript to clarify ALPGM's approach when dealing with the topographical predictors. In a nutshell, ALPGM already uses the technique (b) suggested by the reviewer, but only for the "glacier evolution" component: it learns from interpolated topographical data, as predictors are not necessarily available at each time point, but it is able to simulate glacier-wide SMBs iteratively year-by-year by propagating a newly computed glacier-wide SMB in the model to generate updated topographical predictors for the next year. In more details:

1. In the "glacier-wide SMB modelling" component, which works totally independently from the "glacier evolution" component, the glacier-wide SMB machine learning models are trained based on the historical data available. Topographical and climate data for every year and glacier available in the region of interest is collected. As shown in our case study of French alpine glaciers, the climate data comes from the SAFRAN reanalysis, which is computed for each glacier at its centroid, and the topographical predictors come from the 1967, 1985, 2003 and 2015 glacier inventories (Gardent et al., 2014, with 2015 update). In order to have topographical data for each year, the topographical variables are linearly interpolated between inventories. This is indeed an

approximation and a hypothesis, but as we show throughout the extensive cross-validation, such approximation is enough for the model to understand the relationship between these variables and the glacier-wide SMB. With all this data, the machine learning glacier-wide SMB models are trained and cross-validated using LOGO, LOYO and LSYGO. Therefore, glacier-wide SMB machine learning models are trained with all the training data at once, and then tested on the residual test data for each of the cross-validation folds. After cross-validation, a SMB model is chosen for the spatial, temporal and spatiotemporal dimensions and it is stored as a file which will be later called in the "glacier evolution" component.

2. In the "glacier evolution" component, responsible for coupling the simulated glacier-wide SMBs and the geometry update, these topographical variables are computed differently in order to allow the simulations to be carried out for time periods and glaciers outside the historically observed ones. For each glacier in the region to be simulated, Farinotti et al. (2019) ice thickness and DEM glacier-specific rasters are retrieved for the starting date of the simulation (2003 in our study case). Then, in a loop, for every glacier and year, the topographical predictors are computed from these raster files. Then, the climate predictors at the glacier's current centroid are retrieved from the climate data (e.g. reanalysis or projections) and with all this data the input topo-climatic data for the glacier-wide SMB is assembled. Then, the glacier-wide SMB for this glacier and year is simulated, which combined with the glacier-specific Δh function allows to update the glacier's ice thickness and DEM rasters. This process is repeated in a loop, therefore updating the glacier's geometry with an annual timestep and taking into account the glacier's morphological and topographical changes in the glacier-wide SMB simulations. For the simulation of the following year's SMB, the previously updated ice thickness and DEM rasters are used to re-calculate the topographical parameters, which in turn are used as input topographical predictors for the glacier-wide SMB machine learning model.

This iterative process of re-calculation of topographical predictors before the simulation of the glacier-wide SMB for every year is done in the glacier evolution component, which is intended to be used for glaciers and time periods from the year 2003 onwards. Indeed, since the ice thickness database is based on this year, full simulations with the coupling of the glacier-wide SMB component and the glacier evolution component can only be run from this year onwards. This constraint is however not important for the goal of our general project, which is to be able to simulate the future evolution of French alpine glaciers.

Indeed, few regions in the world have the wealth of data available in the French Alps, especially regarding the multitemporal glacier inventories. Nonetheless, these could be obtained differently in other regions, using global or multiple DEMs covering the region of interest. Moreover, as it is discussed in GC2, the glacier-wide SMB model has proved to still work without any topographical predictors. This aspect is discussed in detail in our reply to GC2.

Regarding Fabien Maussion's comment on the predictor selection procedure, it is discussed in detail in the Specific comments from P8 L1 and P16 L2.

In order to increase the clarity on how the topographical predictors are computed and their different sources, the following changes have been made:

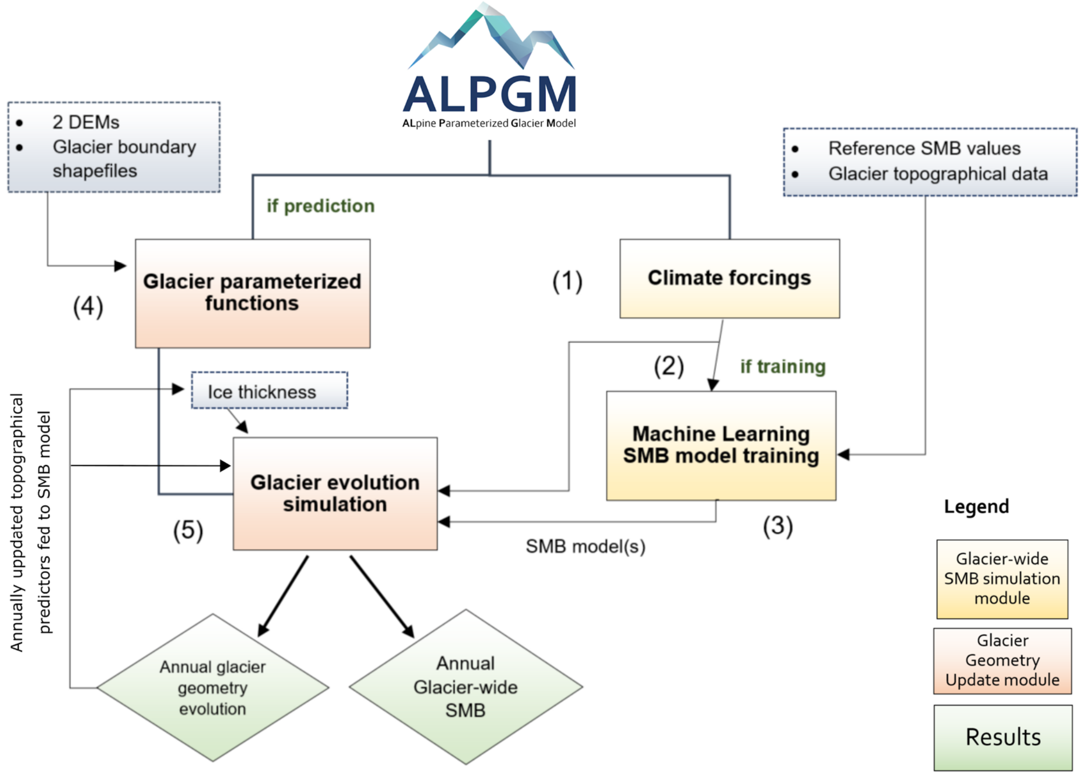
- In Sect. 2.1 "Model overview and workflow":

"2. The SMB machine learning component retrieves the preprocessed **climate predictors from the stored files, retrieves the topographical predictors from the multitemporal glacier inventories**, and then it assembles the training dataset by combining all the necessary topo-climatic predictors. A machine learning algorithm is chosen for the SMB model, which can be loaded from a previous run or it can be trained again with a new dataset. **Then, the SMB model(s) are trained with the full topo-climatic dataset**. These model(s) are stored in intermediate files, allowing to skip this step for future runs."

[..]

"5. Once all previous steps have been run, the **glacier evolution** simulations are launched. For each glacier, the initial ice thickness **and DEM rasters and the glacier geometry update** function are retrieved. **Then, in a loop, for every glacier and year, the topographical data is computed from these raster files. The climate predictors at the glacier's current centroid are retrieved from the climate data (e.g. reanalysis or projections) and with all this data the input topo-climatic data for the glacier-wide SMB model is assembled. Afterwards, the glacier-wide SMB for this glacier and year is simulated, which combined with the glacier-specific geometry update function allows to update the glacier's ice thickness and DEM rasters. This process is repeated in a loop, therefore updating the glacier's geometry with an annual timestep and taking into account the glacier's morphological and topographical changes in the glacier-wide SMB simulations. For the simulation of the following year's SMB, the previously updated ice thickness and DEM rasters is used to re-calculate the topographical parameters, which in turn are used as input topographical predictors for the glacier-wide SMB machine learning model.** If all the ice thickness raster pixels of a glacier become zero, the glacier is considered as disappeared and is removed from the simulation pipeline. For each year, multiple results are stored in data files as well as the raster DEM and ice thickness values for each glacier."

- Fig.1 in Sect. 2.1 "Model overview and workflow", has been updated to indicate the fact that the annually updated topographical predictors are used as input for the glacier-wide SMB model:


ALPGM
ALpine Parameterized Glacier Model
• 2 DEMs
• Glacier boundary shapefiles
if prediction
• Reference SMB values
• Glacier topographical data
Glacier parameterized functions
(4)
(1)
Climate forcings
(2)
if training
Ice thickness
Machine Learning SMB model training
Glacier evolution simulation
(5)
SMB model(s)
(3)
Annually uppdated topographical predictors fed to SMB model
Annual glacier geometry evolution
Annual Glacier-wide SMB
Legend
Glacier-wide SMB simulation module
Glacier Geometry Update module
Results

- In Sect. 3.1.2 "Topographical glacier data and altimetry", the following changes have been added to clarify the source of topographical data at different steps of the model:

"**The topographical data used for the training of the glacier-wide SMB machine learning models is taken from the multitemporal inventory of the French Alps glaciers** (e.g., Gardent et al., 2014) partly available through the GLIMS Glacier Database (NSIDC, 2005). We worked with the 1967, 1985, 2003 and 2015 inventories (Gardent et al., 2014, with 2015 update). **Between these dates, the topographical predictors are linearly interpolated. On the other hand, in the glacier evolution component of ALPGM (Fig.1, step 5), this topographical data is re-computed every year for each glacier from the evolving and annually updated glacier-specific ice thickness and DEM rasters (Sect. 3.1.3). Since these raster files are estimates for the year 2003 (Farinotti et al., 2019 for the ice thickness), the full glacier evolution simulations can start the earliest at this date. For the computation of the glacier-specific geometry update functions, two** DEMs covering the whole French Alps have been used: (1) one from 2011 generated from SPOT5 stereo-pair images, acquired on 15 October 2011; and (2) a 1979 aerial photogrammetric DEM from the French National Geographic Institute (Institut Géographique National, IGN), processed from aerial photographs taken around 1979. Both DEMs have an accuracy between 1 and 4 meters (Rabatel et al., 2016), and their uncertainties are negligible compared to many other parameters in this study."

- Finally, in order to prove that the model, combining the glacier-wide SMB model and glacier evolution components, can successfully simulate the evolution of the topographical parameters and their feedback we have performed a specific test for this. From the year 2003 until 2014, we have run the glacier-wide SMB simulations for the 32 case study glaciers, first with the topographical predictors coming from the interpolated multitemporal glacier inventories (used during the training of the SMB model), and then with the full glacier evolution model, with the Farinotti et al. (2019) ice thickness and DEM raster files. When comparing the results of these simulations, their differences are minimal, mostly coming from the differences between input data (higher resolution DEMs and data from the glacier inventories vs. lower resolution data from Farinotti et al. 2019). For the 2003-2014 period, we obtained very similar performances: an average RMSE of 0.49 m.w.e. a^{-1} using the multitemporal glacier inventories and an average RMSE of 0.52 m.w.e. a^{-1} using the full glacier evolution model with the Farinotti et al. (2019) raster data.

This test, its results and the discussion elements have been added as a new section in the supplementary material, plus a new figure (S6):

3. Topographical glacier-wide SMB predictors

Since topography plays a role in the glacier-wide SMB signal, besides the climate, the representation of the glacier's topography is important in order to correctly simulate its glacier-wide SMB and its geometrical evolution. As explained in Sect. 2.1 "Model overview and workflow" and Sect. 3.1.2 "Topographical glacier data and altimetry", the source of the topographical predictors used for the simulation of glacier-wide SMB is different at different steps of the glacier evolution simulation chain. Two cases exist:

1. For the machine learning training of the glacier-wide SMB models, which is performed on historical data, all topographical data comes from the multitemporal glacier inventories (Gardent et al., 2014, with 2015 update). In order to have an annual timestep, topographical data from these inventories are linearly interpolated.

2. For the full glacier evolution simulation, coupling the glacier-wide SMB component with the glacier geometry evolution component, the model must be capable of generating all the input topographical predictors even for non-observed glaciers and future periods. For every glacier and year, all the topographical predictors are computed from the updated glacier-specific ice thickness and DEM raster files from Farinotti et al. (2019), which then are used to simulate a single glacier-wide SMB for that glacier and year. Then, this glacier-wide SMB together with the glacier-specific geometry update function are used to update the glacier's geometry and their respective ice thickness and DEM rasters. For the next year, all the topographical predictors are recomputed with the updated raster files, and this process is repeated in a loop with an annual timestep. Therefore, the glacier-wide SMB model is called with an annual timestep, simulating only single values in order to take into account the evolution of the glacier's topography.

In order to show that the glacier geometry update component, coupled with the glacier-wide SMB simulation component can successfully simulate the evolution of the topographical characteristics of glaciers in the region, a specific test was designed. Using the same validation period as in Sect. 3.2 (2003-2015), we ran parallel simulations of glacier-wide SMB for all the 32 case study glaciers. The first simulation was done using case (1) with the multitemporal glacier inventories data, and the second one was done following case (2) with the full glacier evolution model and the Farinotti et al. (2019) raster files. The results of both simulations were really similar, revealing only small differences. On average, the simulated glacier-wide SMBs for this period differed on 0.069 m w.e. a^{-1}, due to the differences in the input topographical predictors, which are computed from different datasets (Fig. S6). Moreover, the performances of both simulations for this period are very similar, with a RMSE of 0.49 m.w.e. a^{-1} for case (1) and 0.52 m.w.e. a^{-1} for case (2). The results with all the differences between the simulated glacier-wide SMB values and input topographical values are summarized in Table S1:

Variable (multitemporal inventories vs. full glacier evolution)	**SMB simulated**	**Slope**	**Average glacier elevation**	**Area**
Mean difference	0.069 m.w.e a^{-1}	1.8°	31.3 m	0.2 km^2

Table S1: Differences on simulated glacier-wide SMB and topographical predictors between a simulation using interpolated topographical predictors from the multitemporal glacier inventories and the full glacier evolution simulations including the coupling of the glacier-wide SMB with the glacier geometry update.

The only striking difference is perhaps the difference in simulated areas. This is mainly due to the fact that the Farinotti et al. (2019) dataset uses the RGI v6, which for the largest glaciers of Argentière and Mer de Glace, overestimates its surface area (from 32 to 34 km^2 for Mer de Glace in 2003). The differences in slope are explained by the fact that this variable is not included in the multitemporal glacier inventories (Gardent et al., 2014), therefore it has been computed once with a global DEM and kept constant for each glacier throughout the years for the training of the SMB model. On the other hand, in order to include the long term effects of glacier morphology changes in the glacier evolution simulations (glacier-wide SMB simulation + glacier geometry update), the glacier slope is re-computed with an annual timestep and it evolves through time. Therefore, there are small differences for certain glaciers whose slope has evolved during this period, thus accounting for the differences with the fixed value used for the training of the SMB model.

This test serves to prove that the full glacier evolution simulations in ALPGM are capable of reproducing the topographical predictors used for the training of the glacier-wide SMB machine learning models. Moreover, this test also helps to prove that ALPGM can correctly simulate the topographical evolution of glaciers, which allows to capture the topography induced feedback, which plays a role in the simulation of glacier-wide SMBs.

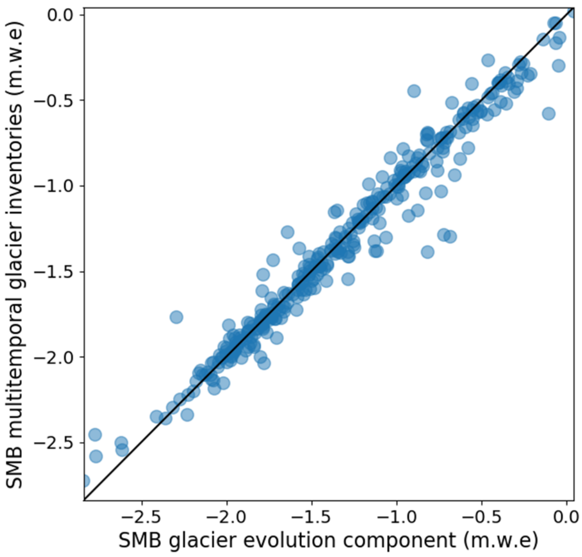


Figure S6: Comparison of glacier-wide SMB simulations (2003-2015, 32 case study glaciers) using topographical predictors from the multitemporal glacier inventories (Y axis) vs. using the full glacier evolution simulations in ALPGM with the Farinotti et al. (2019) ice thickness and DEM rasters (X axis). Average difference = 0.069 m.w.e. a^{-1}

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 xed-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.

This comment raises interesting questions on many levels. As the reviewer mentions, the discussion of glacier-wide *vs.* point/altitude-dependent SMB is a widely discussed topic in the glaciology community. Indeed, glacier-wide SMBs contain not only information on climate but also on the glacier's surface area and topography. Therefore, if a statistical model attempts to simulate glacier-wide SMBs based solely on climate data, it will be missing some information. How important this information actually is, is the main topic of debate, addressed in the two comments from Leclercq et al. and Huss et al. on Huss et al. (2010), as well as in Huss et al. (2012) "Conventional versus reference-surface mass balance".

In most glacier evolution models, SMBs are simulated at different altitudes in order to integrate them in the spatialized modelling framework which takes into account glacier dynamics and ice flow. Nonetheless, this is not necessary when using a glacier geometry update parameterization like the Δh parameterization, as only the total glacier mass annual variation is needed to be redistributed by the Δh function. Therefore, here we chose to work with glacier-wide SMB data due to the fact that in the French Alps there is much more glacier-wide SMB data available (Rabatel et al. 2016) than altitude-dependent or point SMB, essential for training machine learning models. In order to correctly simulate glacier-wide SMB, we included topographical variables that intervene in determining the glacier-wide SMB of glaciers in the French Alps, based on a literature review and on a sensitivity statistical analysis (Sect. "3.2.1 Selection of predictors"). Consequently, our glacier-wide SMB machine learning models do take into account these topographical parameters.

In order to verify these concerns raised by the reviewer, we remade from scratch the causal analysis using Lasso, but this time, no combination of topo-climatic predictors was used. We believe that using linear combinations of topo-climatic predictors increases a lot the number of input predictors, which in turn reduces their relative weight, making the causal analysis and the statistical inference more difficult. Moreover, since the ANN does not take topo-climatic combinations of predictors as input, this new subset of predictors used for the causal analysis is more representative, despite being linear, of the relationships the ANN must be using internally. In this new analysis, the results are similar to the previous causal analysis (Fig. 5). Climate predictors still appear, by far, as the most relevant predictors, with accumulation-related predictors having more importance than ablation-related predictors. Topographical predictors appear to have a minor importance, but they do play a role. Combined, they account for around 3.5% of importance, including the slope of the lowermost 20% altitudinal range, latitude, longitude, area, mean glacier altitude and aspect (Fig. 5). These results are coherent with the spatiotemporal analysis of the glacier-wide SMB signal in the European Alps from Vincent et al. (2017) and Huss (2012), as well as the results from our block cross-validation. The temporal variability of the glacier-wide SMB signal is determined by climate, being the most complex dimension, and topography (differences among glaciers) is modulated by topographical predictors. Therefore, topographical predictors act as spatial modulators, whereas climate acts as the main signal driver, determining the interannual variability. On top of that, as discussed in Sect. "4.3 Perspectives on future applications of deep learning in glaciology", a LOGO cross-validation was done using no topographical predictors, and it was seen that it had a negative impact on the results, with both the accuracy (RMSE) and explained variance (r^2) being negatively affected. This performance penalty (from an RMSE of 0.51 to 0.59 m.w.e. a^{-1}), is indeed small, but it is coherent with the importance attributed to topographical predictors in the Lasso causal analysis (Fig.5).

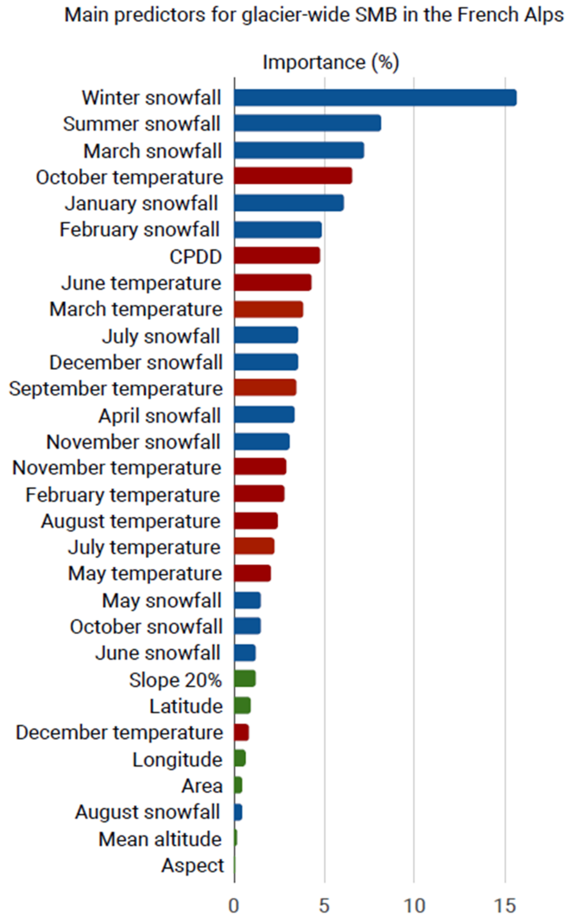


Figure 5. Contribution to the total variance of the 30 top topo-climatic predictors out of 55 predictors using Lasso. Green bars indicate predictors including topographical features, blue ones including accumulation-related features, and red ones including ablation-related features

Therefore, we believe that:

(1) In order to correctly simulate glacier-wide SMBs, topographical predictors play a relevant, albeit secondary, role. They help modulate the glacier-wide SMB signal, introducing spatial changes to the signal determined by climate predictors.
(2) Despite not being possible to "observe" the chosen internal weights and combinations in the ANN, both the Lasso causal analysis and an empirical test removing the input topographical predictors show some benefits in using them. Therefore, it is plausible to conclude that they do play a role in the simulation of glacier-wide SMBs in our case study with French alpine glaciers.

In order to include all these elements of discussion in the manuscript, the following changes have been added:

- **Figure 5** has been updated with the new Lasso causal analysis results.
- The whole Sect. "3.2.2 Causal analysis" has been rewritten to include all these aspects:

"By running the Lasso algorithm on the dataset based on **Eq. 2 and 3**, we obtain the contribution of each predictor in order to explain the annual glacier-wide SMB variance. **Regarding the climatic variables, accumulation-related predictors (winter snowfall, summer snowfall as well as several winter, spring and even summer months) appear as the most important predictors. Ablation-related predictors also seem to be relevant, mainly with CPDD and summer and months at the transition between the seasons (Fig. 5).** Interestingly, meteorological conditions in the transition months are crucial for the annual glacier-wide SMB in the French Alps: (1) October temperature is determinant for the transition between the ablation and the accumulation season, favouring a lengthening of melting when temperature remains positive, or conversely allowing snowfalls that protect the ice and contribute to the accumulation when temperatures are negative; (2) March snowfall has a similar effect: positive anomalies contribute to the total accumulation at the glacier surface, and a thicker snow pack will delay the snow/ice transition during the ablation season leading to a less negative ablation rate (e.g. Fig. 6b, Réveillet et al., 2018). Therefore, meteorological conditions of these transition months seem to strongly impact the annual glacier-wide SMB variability, since their variability oscillates between positive and negative values, unlike the months in the heart of summer or winter.

On the other hand, topographical predictors do play a role, albeit a secondary one. The slope of the 20% lowermost altitudinal range, the glacier area, the glacier mean altitude and aspect help to modulate the glacier-wide SMB signal which partially depends on glacier topography (Huss et al., 2012) unlike point or altitude-dependent SMB. Moreover, latitude and longitude are among the most relevant topographical predictors, which for this case study are likely to be used as bias correctors of precipitation of the SAFRAN climate reanalysis. SAFRAN is suspected of having a precipitation bias, with higher uncertainties for high altitude precipitation (Vionnet et al., 2016). Since the French Alps present an altitudinal gradient, with higher altitudes towards the eastern and the northern massifs, we found that the coefficients linked to latitude and longitude enhanced glacier-wide SMBs with a northeastern gradient."

- The discussion elements in Sect. "4.3 Perspectives on future applications of deep learning in glaciology" have been updated:

"**For ALPGM, the SMB machine learning models were trained using glacier-wide SMB data, due to the high availability of glacier-wide SMB data in the French Alps (Rabatel**

et al., 2016). Nevertheless, the same approach could be used for point SMB data from field observations."

[...]

"In this work, we also evaluated the resilience of the deep learning approach: since many glacierized regions in the world do not have the same amount of data used in this study, we trained an ANN only with monthly average temperature and snowfall, without any topographical predictors, to see until which point the algorithm is capable of learning from minimal data. The results were quite interesting, with a coefficient of determination of 0.68 (against 0.76 from the full model) and a RMSE of 0.59 (against 0.51 from the full model). **These results indicate that meteorological data is the primary source of information, determining the interannual variability of the glacier-wide SMB signal. On the other hand, the "bonus" of topographical data helps to modulate the climate signal, by adding a spatial component to better differentiate glaciers and the topographical characteristics included in the glacier-wide SMB data (Huss et al., 2012).**"

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 h 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).

As explained in our reply to "1 General comments", we made the choice to focus the manuscript on presenting the machine learning glacier-wide SMB modelling approach, combined with the parameterized glacier geometry update, rather than presenting a ready-to-use glacier model. Nonetheless, due to some concerns raised in GC1 and GC2, the focus of the model and the manuscript was likely not clear enough.Through the changes proposed in GC1, GC2 and in this section, we hope to improve the clarity, focus and scope of the paper.

We believe that an important source of confusion might come from the fact that ALPGM is actually, in its current design, a regional glacier evolution model that needs to be trained and applied at a regional scale. This information was nonetheless presented in the manuscript, in the introduction of the machine learning glacier-wide SMB modelling component: since the glacier-wide SMB model needs to be trained with a dataset, the statistical relationships found in datasets are strongest at regional scale, where climate and the glacier's sensitivity to climate remain relatively similar. A more global or continental SMB model could be trained if enough data was available, but nowadays we believe that such an approach is feasible at regional scale only.

- Therefore, ALPGM is now referred to as an "**open-source parameterized regional glacier evolution model**" throughout the manuscript. That is quite a long statement, but at least it is precise.
- In order to set the focus on the approach rather than the presentation of a glacier evolution model, the abstract has been rephrased as follows:

"Abstract. **We present a novel approach to simulate and reconstruct annual glacier-wide surface mass balance series based on a deep artificial neural network (i.e. deep learning). This method has been included as the SMB component of an open-source regional glacier evolution model.** While most glacier models tend to incorporate more and more physical processes, here we take an alternative approach by creating a parameterized model based on data science. Annual glacier-wide SMBs can be simulated **from topo-climatic predictors** using either deep learning or Lasso (regularized multilinear regression), whereas the glacier geometry is updated using a glacier-specific parameterization. We compare and cross-validate our nonlinear deep learning SMB model against other standard linear statistical methods on a dataset of 32 French alpine glaciers. Deep learning is found to outperform linear methods, with improved explained variance (up to +64% in space and +108% in time) and accuracy (up to +47% in space and +58% in time), resulting in an estimated r^2 of 0.77 and RMSE of 0.51 m.w.e. Substantial nonlinear structures are captured by deep learning, with around 35% of nonlinear behaviour in the temporal dimension. For the glacier geometry evolution, the main uncertainties come from the ice thickness data used to initialize the model. These results should encourage the use of deep learning in glacier modelling as a powerful nonlinear tool, capable of capturing the nonlinearities of the climate and glacier systems, that can serve to reconstruct or simulate SMB time series for individual glaciers **in a whole region** for past and future climates."

- The same previous changes in the abstract have been reflected in the conclusions:

"**We presented a novel approach to simulate and reconstruct glacier-wide SMB series using deep learning for individual glaciers at regional scale. This method has been included as a SMB component in ALPGM (Bolibar, 2019), a parameterized regional glacier evolution model, following an alternative approach to most physical and process-based glacier models. The data-driven glacier-wide SMB modelling component is coupled with a glacier geometry update component, based on glacier-specific parameterized functions.** Deep learning is shown to outperform linear methods for the simulation of glacier-wide SMB with a case study of French alpine glaciers. By means of cross-validation, we demonstrated how important nonlinear structures (up to 35%) coming from the glacier and climate systems in both the spatial and temporal dimensions are captured by the deep ANN. Taking into account this nonlinearity substantially improved the explained variance and accuracy compared to linear statistical models, especially in the more complex temporal dimension. As we have shown in our case study, deep ANNs are capable of dealing with relatively small datasets, and they present a wide range of configurations to generalize and prevent overfitting. Machine learning models benefit from the increasing number of available data, which makes their performance constantly improve as time goes by."

Regarding point 3, we believe that the glacier geometry update component as well as its section in the manuscript should be kept. Indeed, the main novelty of the study and the model is its glacier-wide SMB component, but the fact that there is a glacier geometry update component makes it possible to simulate the evolution of glaciers at regional scale. We agree with the reviewer that the main focus of the study should be points 1 and 2, but we believe that point 3 can still be kept. With the changes presented here, we intend to give the main focus of the paper on the novel machine learning glacier-wide SMB approach, and mention the regional glacier evolution model as a platform in which this method has been implemented as a SMB component. It is important to present the combined SMB and glacier geometry update components, since this is a methodology paper, which will serve as a reference in the two future results papers. With the clarifications made in GC1 and GC2, and the fact that we specify that ALPGM is a regional model, we believe that the statement that ALPGM is a glacier evolution model is fair.

2 Specific comments

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

The LOYO cross-validation evaluates the model's performance for time periods outside the data used for training. Since cross-validation can only be performed with past observed data, we believe that the current analysis from this study allows us to determine that the SMB modelling approach can be used in unseen (past or future) periods or climates. The only way to demonstrate that a model can be used for future climates is to validate it in the past over unseen periods (using LOYO), and to anticipate possible issues linked to hypotheses that might not be fulfilled.

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.

We agree that the current model classification is rather done based on glacier dynamics than SMB modelling. These two main components of glacier models have been identified in this section, but then they are mixed in the classification. Nonetheless, since ALPGM is a glacier evolution model and not just a SMB model, we believe that it is better to create two separate lists, one for SMB modelling and another one for glacier dynamics, instead of removing the glacier dynamics part which concerns the Δh parameterization used in ALPGM to update the glacier geometry.

Therefore, this section of the "Introduction" has been updated as follows:

"Glacier and hydro-glaciological models can help answer these questions, giving several possible outcomes depending on multiple climate scenarios. **(a) Surface mass balance (SMB) and (b) glacier dynamics both need to be modelled to understand glacier evolution on regional and sub-regional scales. Models of varying complexity exist for both processes.** In order to model these processes at large scale (i.e. on several glaciers at a catchment scale), some compromises need to be made, which can be approached in different ways:

(a) Regarding SMB:

1. Empirical models, like the temperature-index model (e.g. Hock, 2003), simulate glacier SMB through empirical relationships between air temperature and melt and snow accumulation.

2. Statistical or machine learning models describe and predict glacier SMB based on statistical relationships found in data from a selection of topographical and climate predictors (e.g. Martin, 1974; Steiner et al., 2005).

3. Physical and Surface Energy Balance (SEB) models take into account all energy exchanges between the glacier and the atmosphere, and can simulate the spatial and temporal variability of snowmelt and the changes in albedo (e.g. Gerbaux et al., 2005).

(b) Regarding glacier dynamics:

1. Parameterized models do not explicitly resolve any physical processes, but implicitly take them into account using parameterizations, based on statistical or empirical relationships, in order to modify the glacier geometry. This type of models range from very simple statistical models (e.g. Carlson et al., 2014) to more complex ones based on different approaches, such as a calibrated equilibrium-line altitude (ELA) model (e.g. Zemp et al., 2006), a glacier retreat parameterization specific for glacier size groups (e.g. Huss and Hock, 2015) or volume/length-area scaling (e.g. Marzeion et al., 2012; Radic et al., 2014).

2. Process-based models, like GloGEMflow (e.g. Zekollari et al., 2019) and OGGM (e.g. Maussion et al., 2019), approximate a number of glacier physical processes involved in ice flow dynamics using the shallow ice approximation.

3. Physics-based models, like the finite elements Elmer/Ice model (e.g. Gagliardini et al., 2013), approach glacier dynamics by explicitly simulating

> **physical processes and solving the full Stokes equations (e.g. Jouvet et al., 2009; Réveillet et al., 2015)..**

At the same time, **the use of** these different approaches strongly depend on available data, whose spatial and temporal resolutions have an important impact on the results' quality and uncertainties (e.g., Réveillet et al., 2018). **Parameterized glacier dynamics models and empirical and statistical SMB models** require a reference or training dataset to calibrate the relationships, which can then be used for projections with the hypothesis that relationships remain stationary **in space or in time**. On the contrary, process-based and specially physics-based **glacier dynamics and SMB models** have the advantage of representing physical processes, but they require larger datasets at higher spatial and temporal resolutions with a consequently higher computational cost (Réveillet et al., 2018). **For SMB modelling, meteorological** reanalyses provide an attractive alternative to sparse point observations, although their spatial resolution and suitability to complex high-mountain topography are often not good enough for high-resolution physics-based glacio-hydrological applications. However, parameterized models are much more flexible, equally dealing with fewer and coarser meteorological data as well as the state of the art reanalyses, which allows to work at resolutions much closer to glaciers' scale and to reduce uncertainties. The current resolution of climate projections is still too low to adequately drive most glacier physical processes, but the ever-growing datasets of historical data are paving the way for the training of parameterized machine learning models."

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

Mainly oceanography (e.g. Ducournau and Fablet, 2016; Lguensat et al., 2018) and climatology (e.g. Rasp et al., 2018; Jiang et al., 2018). The large datasets of ocean remote sensing data and climate reanalysis are being treated with approaches mixing deep learning, machine learning and physics. The scientific communities are much more advanced towards data-driven approaches, and have started to bridge the gap between data scientists and geophysicists, by adding physical knowledge into data science models.

These aspects have been added to the manuscript:

"Compared to other fields in geosciences**, such as oceanography (e.g., Ducournau and Fablet, 2016; Lguensat et al., 2018), climatology (e.g., Rasp et al., 2018; Jiang et al., 2018) and hydrology (e.g. Marçais and de Dreuzy, 2017; Shen, 2018),** we believe that the glaciological community has not yet exploited the full capabilities of these approaches."

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.

As suggested by both reviewers, this sentence has been rephrased in order to remove any negative or subjective connotations:

"… **we believe that the glaciological community has not yet exploited the full capabilities of these approaches**."

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 3 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 nonlinear. 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.

Indeed, here we refer to a statistical linearity. The example given by Fabien Maussion (with a temperature-index model being non-linear) would therefore be comparable to the simple multiple linear regression used in many studies, fed by PDDs and cumulative snowfall. By "pre-processing" the input predictors (daily temperatures to CPDD and precipitation to cumulative snowfall) one is introducing nonlinearities in the system. Nonetheless, this just creates new predictors which in this new space, can also behave linearly or nonlinearly. Indeed, the relationship between temperature and melt in the models you mention is nonlinear, but between PDDs and melt it is linear. So what we are referring to here, is that no matter how the input physical variables are pre-processed, their new relationships in the transformed space will probably be nonlinear as well. Therefore, there should be some benefits (as shown in this study) by switching to nonlinear models, even with the pre-processing of input predictors.

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).

This is related to the aspects raised in the P2 L6-16 comment. We are referring to both SMB and ice dynamics. For GloGEMflow, the addition of ice dynamics with the shallow ice approximation, and for OGGM, the ongoing developments on the SMB model (with the addition of shortwave radiation for instance). ALPGM does not use physics neither in SMB nor ice dynamics modelling. Therefore, the sentence has been updated as follows:

"When most glacier evolution models tend to incorporate more and more physical processes **in SMB or ice dynamics** (e.g., Maussion et al. (2019); Zekollari et al. (2019)), ALPGM takes an alternative approach based on data science **for SMB and parameterizations for glacier dynamics**."

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

Here we intended to make a quick introduction to the SMB modelling approach, and more specifically say that our approach aims at being a regional approach, which requires to be trained for each region of interest.

As suggested by the reviewer, this paragraph has been moved to the discussion section, in Sect. 4.2, at the end. Moreover, the paragraph from P5 L29-34 has been modified accordingly, to avoid any discussion items and just state the fact that the SMB modelling approach is regional:

"Annual glacier-wide SMBs are simulated using machine learning. **Due to the regional characteristics and specificities of topographical and climate data, this glacier-wide SMB modelling method is, for now, a regional approach.**"

P6 L25: StatsModel" is spelled "StatsModels"

This has been updated as suggested by the reviewer.

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.

This sentence is in fact deprecated and it is now removed from the manuscript. This is also related to the comments in the last paragraph of GC1. During the development of the study, predictor selection via the Lasso was a hypothesis to be tested, but empirical tests using subsets of topo-climatic predictors as inputs of the ANN showed that it did not improve the results at all. The ANN is capable of choosing the relevant predictors by setting the weights of the connections of non-important predictors to zero. Moreover, as suggested by the reviewer, for the sake of equality in the comparison between statistical methods, the Lasso and ANN are fed with the same topo-climatic predictors.

The choice of input topo-climatic predictors is explained in Sect. 3.2.1; first with a literature review to target potential explanatory variables, and then with individual linear regressions to test the sensitivity of the SMB data to each individual predictor, similarly to what was done in Rabatel et al. (2016). This first choice of predictors is then used by each of the 3 statistical approaches: (1) All-possible multiple regressions tests the performance of all the possible subsets of predictors, (2) the Lasso performs a coefficient shrinkage to regularize the input predictors in order to discard them in a continuous way, and (3) the ANN gives specific weights to each connection of combined and non-combined input predictors at each neuron.

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.

Indeed, the hyperparameter selection needs to be done using cross-validation, but the process is quite different for Lasso than for an ANN.

For Lasso, there is only one hyperparameter (the α value), which is determined using cross-validation. This can be done with the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC) and a classical cross-validation with iterative fitting along a regularization path. This is explained in Sect. 2.2.3 "Lasso". Sci-kit learn provides different classes to help choose the best α value before fitting the Lasso model to data.

For the ANN, this becomes quite more complex. The list of hyperparameters to be fine-tuned is very long, so a brute force strategy of grid search or cross-validation of every single hyperparameter is strongly time consuming. There are smarter ways to proceed, especially with the knowledge of which types of optimizers, activation functions and hyperparameter values work best together. These aspects are discussed in Sect. 4.2. In an early stage, the ANN architecture with the number of neurons and layers, as well as the learning rate and type of optimizer, were cross-validated or tested for a few folds of the data. This early hyperparameter selection is not a proper cross-validation from A to Z, since in order to narrow down the range of values which gives the best performance, tests were first done in a few random folds. Then, once one has a better view of what can work for this dataset and architecture, a full cross-validation is performed with a few candidate hyperparameter values to choose the final ones. Therefore, in some ways it works similarly to Lasso and any machine learning training, where there is a first stage of hyperparameter tuning in order to choose a final configuration. Once the hyperparameters have been chosen, they remain constant throughout all the cross-validation, as it is done with Lasso. Due to the complexity of ANNs, the great number of hyperparameters and the fact that everything is open-source with new optimizers and approaches being released every month, the hyperparameter fine-tuning is a process that could be taken to infinity. One needs to know when to stop, when gains and additional tests stop bringing much added performance to the model.

These aspects have been added to the discussion in Sect. 4.2:

"In order to cope with the specific challenges related to each type of cross-validation, there are several hyperparameters that can be modified to adapt the ANN's behaviour. **Due to the long list of hyperparameters intervening in an ANN, it is not advisable to select them using brute force with a grid search or cross-validation. Instead, initial tests are performed in a subset of random folds to narrow down the range of best performing values, before moving to the full final cross-validations for the final hyperparameter selection.**"

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.

It depends on what we understand by "unseen glaciers". The Δh methodology used in Huss and Hock (2015), with a parameterization specific to 3 glacier sizes, requires to know at least the area of all the glaciers to be simulated, in order to choose a Δh function for them. In our case, in ALPGM, a glacier-specific Δh is computed for each glacier based on past DEMs. As explained in Sect. 3.3, these two DEMs cover the whole region of interest, therefore Δh functions are calculated for each glacier of the French Alps. With the SMB model trained from the 32 glaciers from the case study, the annual glacier-wide SMBs of all the glaciers in the French Alps can be simulated (the results will soon be presented in a separate paper), and since we have a specific Δh function for each glacier, their geometry can be updated with a better accuracy than using the 3 size-specific Δh functions used by Mathias Huss' studies. Therefore, ALPGM is capable of simulating the evolution of "unseen" glaciers (*i.e.* glaciers outside the 32 case study ones) in the same region.

Huss et al. (2008b), first presented the "Δh parameterization" for individual glaciers, as done in ALPGM, and they later presented the size-averaged functions in Huss et al. (2010). We believe that, despite being computed specifically for each glacier, the use of the word "parameterization" is fair in this context. The glacier-specific functions replace unresolved physical processes, and they rely on input data that is available for all glaciers. A parameterization refers to the procedure of replacing complex (geophysical in this case) processes by simplified processes. In the words of Stensrud (2007), "Parameterization schemes": "There are always physical processes and scales of motion that cannot be represented by a numerical model, regardless of the resolution […]. Parameterization is the process by which the important physical processes that cannot be resolved directly by a numerical model are represented".

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).

The results of the proposed glacier-wide SMB reconstruction methodology of this paper applied to all the glaciers in the French Alps will soon be presented in a separate paper. Indeed, it would be a complementary way to cross-validate the glacier-wide SMB model, but the extension to other glaciers outside the ones from the case study is out of the scope of this (already quite long) paper. Moreover, comparing simulated annual glacier-wide SMBs to cumulative goedetical SMBs is quite limited, as it only serves to determine the bias of the model, since simulations would have to be summed in order to produce the interannual glacier-wide SMB, with positive and negative errors potentially being compensated or cancelled among them.

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.

This aspect has been addressed together with the next comment. Please see the updated Sect. 3.1.2 from the next comment.

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).

Sect. 3.1.2 has been updated in order to include this information:

"**The topographical data used for the training of the glacier-wide SMB machine learning models is taken from the multitemporal inventory of the French Alps glaciers (e.g. Gardent et al. (2014)) partly available through the GLIMS Glacier Database (NSIDC (2005)). We worked with the 1967, 1985, 2003 and 2015 inventories (Gardent et al. (2014), with 2015 update). Between these dates, the topographical predictors are linearly interpolated. On the other hand, in the glacier evolution component of ALPGM (Fig.1, step 5), the topographical data are re-computed every year for each glacier from the evolving and annually updated glacier-specific ice thickness and DEM rasters (Sect. 3.1.3). For the computation of the glacier-specific geometry update functions, two DEMs covering the whole French Alps have been used:** (1) one from 2011 generated from SPOT5 stereo-pair images, acquired on 15 October 2011; and (2) a 1979 aerial photogrammetric DEM from the French National Geographic Institute (Institut Géographique National, IGN), processed from aerial photographs taken around 1979. Both DEMs have an accuracy between 1 and 4 meters (Rabatel et al. (2016)), and their uncertainties are negligible compared to many other parameters in this study."

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.

The fact that no combination of topo-climatic predictors is done for the ANN (unlike for Lasso as shown in Eq. 4) is explained in Sect. 2.2.4 "Deep artificial neural network", lines 8-10. As explained in our response from comment P8 L1, no predictor selection is done for the ANN. Each of the 3 machine learning algorithms performs its own predictor selection.

In order to aid the reader, this information has been added again in a brief way in this section:

"For the training of the ANN, no combination of topo-climatic predictors is done as previously mentioned **(Sect. 2.2.4), since it is already done internally by the ANN**."

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).

Meteorological reanalyses are not capable of perfectly reproducing all the complex precipitation patterns found in mountainous regions. As explained in Vionnet et al. (2016), the SAFRAN reanalysis is more uncertain, and likely negatively biased at the higher altitudes in the French Alps. Because of the distribution of high altitude massifs towards the northern and eastern sides of the French Alps, this precipitation bias is more present in the northern massifs (see Table 6 from Vionnet et al. 2016 included in this reply) and the eastern massifs.

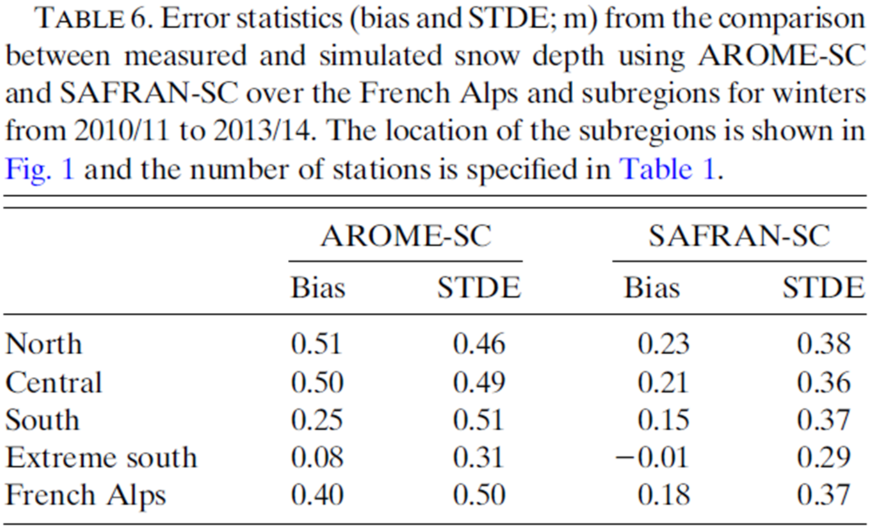


TABLE 6. Error statistics (bias and STDE; m) from the comparison between measured and simulated snow depth using AROME-SC and SAFRAN-SC over the French Alps and subregions for winters from 2010/11 to 2013/14. The location of the subregions is shown in Fig. 1 and the number of stations is specified in Table 1.

	AROME-SC		SAFRAN-SC	
	Bias	STDE	Bias	STDE
North	0.51	0.46	0.23	0.38
Central	0.50	0.49	0.21	0.36
South	0.25	0.51	0.15	0.37
Extreme south	0.08	0.31	−0.01	0.29
French Alps	0.40	0.50	0.18	0.37

We believe, as suggested by the reviewer, that these two topographical parameters play the role of bias correction. Since the machine learning models learn from past data, they can find relationships between precipitation and regions, compensating some of the bias found in the climate data. A quick analysis of the coefficients from the Lasso causal inference analysis revealed that the latitude and longitude predictors modulate the glacier-wide SMB signal in the French Alps, with a positive northeastern gradient. Therefore, these two predictors likely correct the underestimation of precipitation for the northeastern massifs of the French Alps. In the original Lasso causal inference analysis, where the combination of topo-climatic predictors was taken into account, it was shown how longitude*winter snowfall and latitude*winter snowfall were among the most important predictors. This further assesses the importance of latitude and longitude to correct the bias of snowfall in certain regions.

This information has been added in Sect. 3.2.2 “Causal analysis”:

“**In a second term, topographical predictors do play a role, albeit a secondary one. The slope of the 20% lowermost altitudinal range, the glacier area, the glacier mean altitude and aspect help to modulate the glacier-wide SMB signal, which unlike point or altitude-**

dependent SMB, partially depends on glacier topography (Huss et al., 2012). Moreover, latitude and longitude are among the most relevant topographical predictors, which for this case study are likely to be used as bias correctors of precipitation of the SAFRAN climate reanalysis. SAFRAN is suspected of having a precipitation bias, with higher uncertainties for high altitude precipitation (Vionnet et al., 2016). Since the French Alps present an altitudinal gradient, with higher altitudes towards the eastern and the northern massifs, we found that the coefficients linked to latitude and longitude enhanced glacier-wide SMBs with a north-east gradient."

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.

We agree with the reviewer on that point, provided perfect, unbiased climate reanalysis were available. However, as explained in the previous comment, the latitude and longitude topographical predictors play in our case a role in precipitation adjustment/bias correction. The use of these predictors does not hinder the applicability of this approach, since as mentioned in Sect. 3.2, this glacier-wide SMB modelling approach is regional, and the training is specific for each region (determined by the spatial coverage of climate forcings and their quality, the sensitivity of SMB to climate and topographical data). If this SMB modelling approach was to be applied for instance to the whole European Alps, it would require first to switch to climate forcings with a full coverage of the European Alps, add as much SMB data of glaciers in the region as possible and then to retrain the model.

ALPGM should not be mistaken for a global glacier evolution model. The way the SMB machine learning models have been trained so far, is region-based. In the future, with an ambitious training with lots of data, perhaps it would be possible to make the SMB model more global. But it would require re-thinking many things, in order to consider different SMB-climate sensitivities.

As discussed in GC3, the manuscript has been updated, referring to ALPGM as a regional model throughout the different sections.

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.

We agree that the min-max ranges should be the same. However, the fact that linear methods are in purple-red and nonlinear in blue-green is not random. This colour code is respected throughout the paper (Figs. 6, 7, 8, 9 and 10), with all figures referring to the different models. With the same min-max range, since the colour gradient is extremely similar, it is very easy to compare linear and nonlinear plots while keeping the color code. It is worth mentioning that keeping the min-max range constant has improved the figures, allowing them to show how for LOYO, deep learning performs slightly worse than for LOGO.

The plots have been updated, changing the font size of the axis and keeping a constant min-max range for all the scatter plots.

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.

As suggested by the reviewer, we have added an extra plot with the error distribution of the deep learning glacier-wide SMB model, in the temporal and spatial dimensions. The effect explained in these sections shows up, with higher errors for extreme values, mostly due to their underrepresentation in the dataset, being outliers. This new plot has been added in the supplementary material and it has been added as a reference during the discussion involving Fig. 6 and 7.

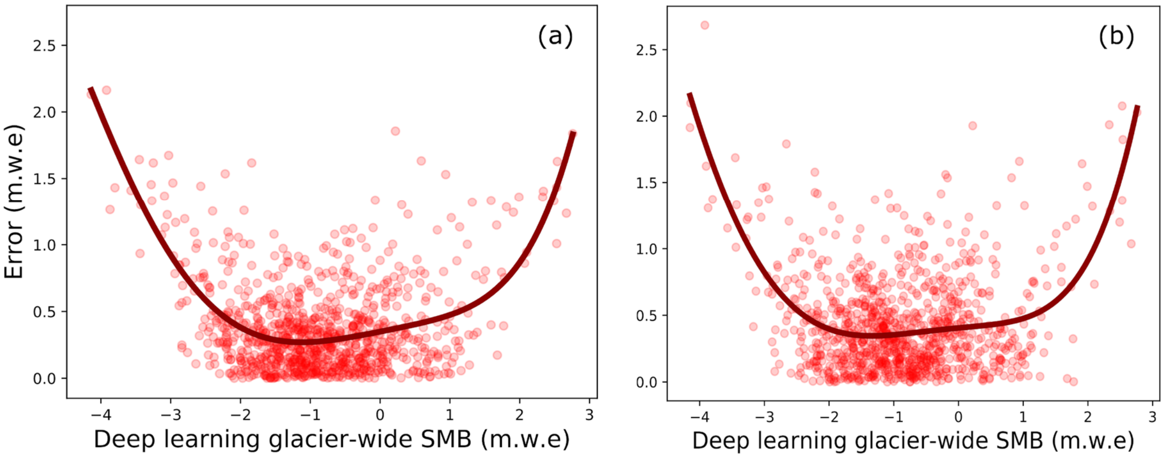
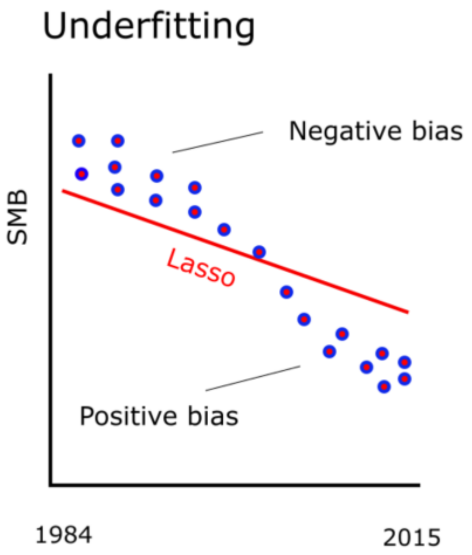


Figure S7: Error distribution of deep learning (without weights) glacier-wide SMB simulations for the 1984-2015 period for the 32 case study glaciers. (a) Performance in the spatial dimension using LOGO cross-validation; (b) performance in the temporal dimension using LOYO cross-validation. The red line corresponds to a 5th order polynomial fit.

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?

This aspect has been briefly discussed in the last sentence in Sect. 3.2.4 "Temporal predictive analysis". The lack of complexity of the linear Lasso is more obvious in the more complex temporal dimension compared to the spatial dimension. As explained in Sect. 4.2 "Training deep learning models with spatiotemporal data", the interannual variability in the glacier-wide SMB signal (temporal dimension) is controlled by climate, and the topography (spatial dimension) is responsible for the modulation of the signal. This lack of complexity of the linear model towards the temporal glacier-wide SMB signal, means that the model is underfitting the data, and by trying to minimize the overall error (increasing the variance), it introduces a different (negative for the first half and positive for the second) bias for the two clearly distinct periods of glacier-wide SMB. This means that the Lasso model manages to have an acceptable RMSE at the price of introducing an important bias.



This explanation has been extended to address these details:

"The important bias present only with Lasso is representative of its lack of complexity towards nonlinear structures, **which results in an underfitting of the data. The average error is not bad, but it shows a high negative bias for the first half of the period, which mostly has slightly negative glacier-wide SMBs, and a high positive bias for the second half of the period, which mostly has very negative glacier-wide SMB values**."

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.

Indeed, it can be seen as a bias test since the errors are accumulated throughout the years, and the final result is the one which is observed and compared. Nonetheless, it serves to test that the glacier evolution model (SMB component + glacier geometry update + ice thickness data) is capable of reproducing past glacier evolution. Since the glacier-wide SMB data used here is cross-validated, this test is representative of the model's end-to-end performance. More details regarding the uncertainties and the purpose of Fig. 11 are discussed in the next question.

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 h method raised above would remain, though.

Indeed, the fact that we said that the main uncertainty came from the initial ice thickness data used was purely based on similar studies (e.g. Huss and Hock, 2015) which claimed this. In order to verify this by ourselves, we have added the uncertainty linked to the Δh glacier geometry function calculation in Fig. 11. From the automatic process to compute it from the DEM difference, we estimated an average uncertainty of ±10% based on the polynomial fits over the altitude difference values (Fig. S4). By adding this uncertainty, we can now claim, as stated by other studies, that the uncertainties linked to the initial ice thickness are greater than the uncertainties linked to the glacier geometry update parameterization.

Regarding the out of sample validation proposed here, we believe that training the model with only data before 2003 would not be representative of the model’s real performance, since we would be using only half of the data available, thus severely penalizing the model. Instead, we have used the LOGO models, which allow a true out of sample validation for each glacier. These results have been included in Fig. 11 as well. Therefore, now the test of Fig. 11 allows to show the end-to-end glacier evolution performance of ALPGM for the 32 case study glaciers, including a true out of sample glacier-wide SMB values, the uncertainties linked to the initial ice thickness data and the uncertainties linked to the glacier geometry update functions. Indeed, there is no way to truly cross-validate these geometry update functions, since they need to be calibrated for the longest available period (1979-2011), but we believe that this test still allows to quantify and prove the glacier evolution simulations’ performance.

The explanations in Sect. 3.3 “Glacier geometry evolution: Validation and results” have been updated accordingly, including Fig. 11:

“In order to evaluate the performance of the parameterized glacier dynamics of ALPGM, **coupled with the glacier-wide SMB component,** we compared the simulated glacier area of the 32 studied glaciers with the observed area in 2015 from the most up-to-date glacier inventory in the French Alps. Simulations were started in 2003, for which we used the F19 ice thickness dataset. In order to take into account the ice thickness uncertainties, we ran three simulations with different versions of the initial ice thickness: the original data, -30% and +30% of the original ice thickness in agreement with the uncertainty estimated by the authors. **Moreover, in order to take into account the uncertainties in the ∆h glacier geometry update function computation, we added a ±10% variation in the parameterized functions** (Fig. 11).

Overall, the results illustrated in Fig. 11 show a good agreement with the observations. Even for a 12-year period, the initial ice thickness remains the largest uncertainty, with almost all glaciers falling within the observed area when taking it into account. The mean error in simulated surface area was of **10.7%** with the original F19 ice thickness dataset. Other studies using the ∆h parameterization already proved that the initial ice thickness is the most important uncertainty in glacier evolution simulations, together with the choice of a GCM for future projections (Huss and Hock, 2015)."

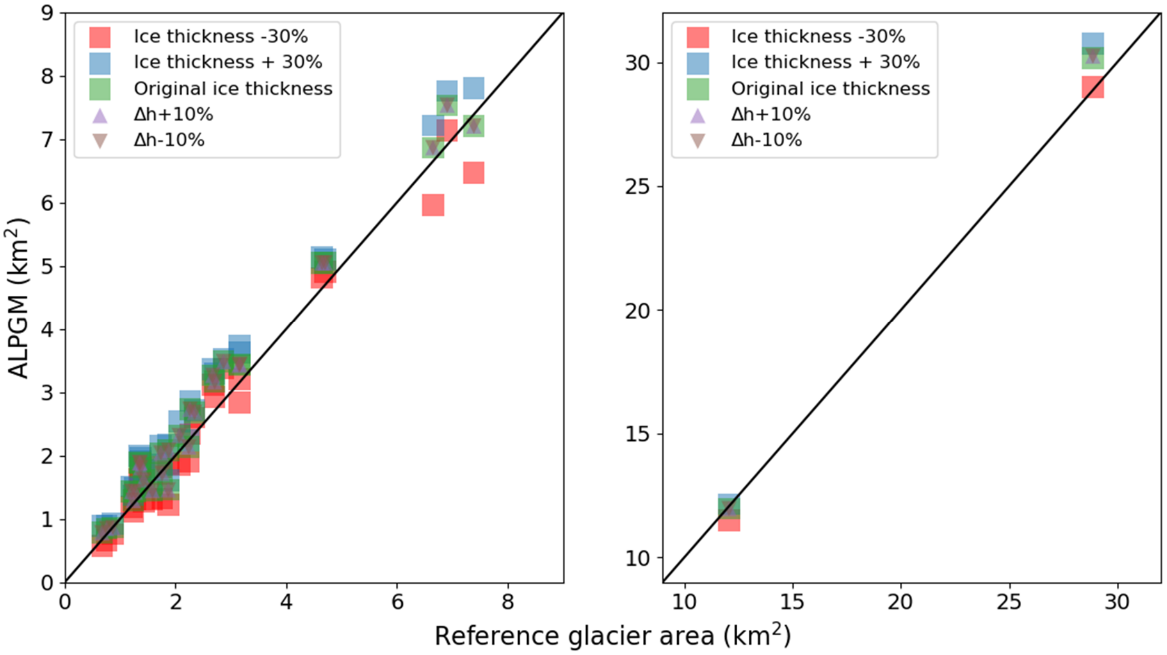


Figure 11. Simulated glacier areas for the 2003-2015 period for the 32 study glaciers using a deep learning SMB model without weights. Squares indicate the different F19 initial ice thicknesses used taking into account their uncertainties and triangles the uncertainties linked to the glacier-specific geometry update functions. For better visualisation, the figure is split in two with the two largest French glaciers on the right.

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).

This aspect is discussed in detail in our replies to GC1 and GC2. The good performance of a glacier-wide SMB model without topographical predictors, shows again the secondary role that they play. As discussed in Huss et al. (2012) "Conventional versus reference-surface mass balance", and it the comments in The Cryosphere between Leclrercq et al. and Huss et al., there is a debate regarding their importance, but from a statistical point of view based on the data used in this study, we can state that in our case they do play a secondary role.

Anonymous Referee #1

The research presented in this manuscript shows promising results in the application of an ANN model used for surface mass balance modelling. The manuscript is, for the most part, well organized. The manuscript can be greatly improved by increasing clarity and specificity throughout. I hope that my comments are helpful to the authors in this effort. No single one of my comments identifies a major flaw with the manuscript; rather, there are many small changes that I believe can be made to improve the quality of the paper. I have organized my comments in sequential order by section, preceded by one general note.

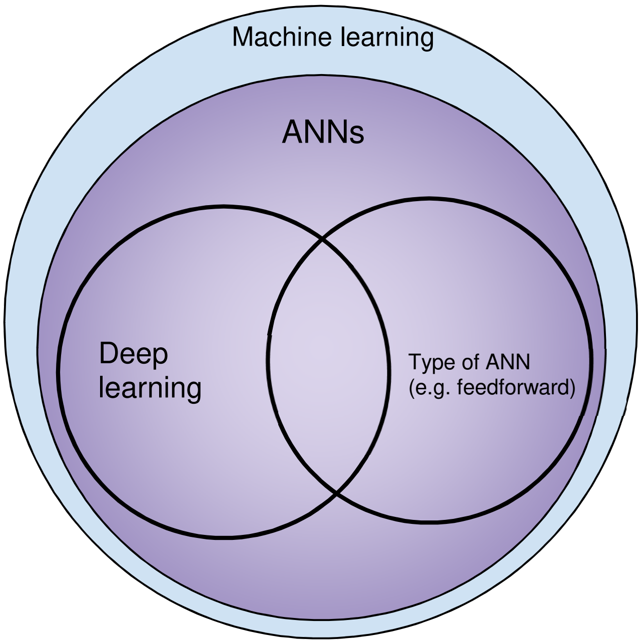
We would like to thank the reviewer for the time dedicated to read the manuscript and for the overall positive feedback. All the detailed comments will hopefully improve the overall clarity and fluidity of the manuscript. All points raised during the review have been addressed and answered, in the following detailed sections, and the manuscript has been updated accordingly.

Small changes within paragraphs are shown in bold, in order to distinguish them from their context.

General Note:

The difference between "machine learning" and "deep learning" is not clearly defined in the literature, but a 6-layer ANN is likely at the very tip of what may constitute "deep learning". Considering that, as you note, deep learning is not a common tool among the glaciological community, it would be good to provide further context as to what an ANN is (a type of model, which is relatively simple in the deep learning world as compared to, say, a convolutional neural network or long short-term memory network) versus what deep learning is. I believe that this is required especially because the title refers to deep learning broadly, not a deep ANN specifically, and it should be made clear that there is much more to deep learning than ANNs.

Indeed, the jargon in the machine and deep learning fields is often not well defined. Nonetheless, deep learning is a subfield of machine learning, involving ANNs with more than one hidden layer. Therefore, one could determine the following hierarchy between these concepts:



ANNs are an example of machine learning, and within ANNs one needs to choose an architecture: single or deep (multiple) hidden layers, and a type of ANN: feedforward (used in our study), convolutional, LSTM.

The title refers to deep learning, which is broad in the sense that it could imply the use of different types of ANN. But in Sect. "2.2.4 Deep artificial neural network" line 16, we specify that a feedforward fully-connected ANN is used. We do not understand what the reviewer means by "and it should be made clear that there is much more to deep learning than ANNs", since deep learning is a subfield of machine learning constituted only by ANNs with multiple hidden layers.

In order to increase clarity for the reader, we have specifically mentioned this aspect in Sect. 2.2.4, in lines 16-22, which now reads as:

"Artificial neural networks (ANNs) are nonlinear statistical models inspired by biological neural networks (Fausett (1994); Hastie et al. (2009)). A neural network is characterized by: (1) the architecture or pattern of connections between units and the number of layers (input, output and hidden layers); (2) the optimizer: the method for determining the weights of the connections between units; and (3) its (normally nonlinear) activation functions (Fausett (1994)). **When ANNs have more than one hidden layer (*e.g.* Fig. 3), they are referred to as deep ANNs or deep learning**. The description of neural networks is beyond the scope of this study, so for more details and a full explanation please refer to Fausett (1994), Hastie et al. (2009), as well as Steiner et al. (2005, 2008) where the reader can find a thorough introduction to the use of ANNs in glaciology."

Authors reply to Anonymous Referee #1's review on "Deep learning applied to glacier evolution modelling"

Sequential notes:

Page 1, Line 22: What does "individual glaciers at regional scale" mean? Do you mean to say, "individual glaciers within the same region"?

We mean the reconstruction of SMB series of individual glaciers for a whole region.

The sentence has been rephrased to improve its clarity:

"… that can serve to reconstruct or simulate SMB time series for individual glaciers in a whole region for past and future climates."

1 Introduction: Page 1, Line 25: ". . . being climate proxies which can clearly depict the evolution of climate for the global audience"; remove "clearly", if the evolution of climate was clear for the global audience, then why is there so much disagreement among the global audience?

The sentence has been adapted as suggested by the reviewer.

Page 1, Line 26: "For the coming decades . . ."; I believe this should be "In the coming decades . . ."

The sentence has been adapted as suggested by the reviewer.

Page 1, Line 28: "The reduction in ice volume may produce an array of consequences which requires to be properly predicted." This sentence, and the following, is vague. What consequences are you talking about? Be explicit.

The sentence has been rephrased as it follows to specify the consequences and importance of glacier retreat:

"**The reduction in ice volume may produce an array of hydrological, ecological and economic consequences in mountain regions which requires to be properly predicted**. These consequences will strongly depend on the future climatic scenarios, which will determine the timing and magnitude for the transition of hydrological regimes (Huss and Hock (2018)). **Understanding these future transitions is key for societies to adapt to future hydrological and climate configurations.**"

Authors reply to Anonymous Referee #1's review on "Deep learning applied to glacier evolution modelling"

Page 2, Line 2: "For any glacier model. . ."; Saying "any" makes this sentence too broad and not necessarily true. Be explicit for the classes/types/purposes of modes which require SMB and glacier dynamics (e.g. "SMB and glacier dynamics both need to be modelled to understand glacier evolution on regional and sub-regional scales. Models of varying complexity exist for both processes.")

The sentence has been rephrased as suggested by the reviewer.

Page 2, Line 18: ". . .these different approaches strongly depend on available data. . ."; Change to ". . .the use of these different approaches strongly depend on available data. . ." since it the model usage, not the model itself, which depends on what data one has.

The sentence has been rephrased as suggested by the reviewer.

Page 2, Line 21: ". . .relationships remain stationary."; Change to ". . .relationships remain stationary in time."

The sentence has been rephrased as follows, including as well the spatial dimension:

"… which can then be used for projections with the hypothesis that relationships remain stationary in time."

Page 2, Line 34: ". . .the glaciological community has remained quite oblivious to these advances. . ."; Oblivious is a strongly negative word to use here, and it is a disservice to insult your readers.

We agree that this word choice is not the most suitable in this context, because of its negative and subjective connotations. The sentence has been rephrased as follows:

"… **we believe that the glaciological community has not yet exploited the full capabilities of these approaches**."

Page 6, Line 10: ". . .relevant predictors must be selected, performing a sensitivity study. . ."; Change to ". . .relevant predictors must be selected, so we perform a sensitivity study. . ."

The sentence has been rephrased as suggested by the reviewer.

Page 6, Line 14 and Equation 1: Is there a reference for this "effective way of expanding the training dataset"?

This is a common practice in regression, similarly to data augmentation and what an ANN does internally combining the input parameters in each hidden layer. It must of course be done before subset selection or regularization. It is explained in Weisberg (2014), Sect. 10.2, which has been added as a reference for this sentence.

Page 7, last sentence: Here you describe the types of cross validations available in ALPGM. Which did you use?

We use the cross-validation with iterative fitting along a regularization path. This has been now specified after the sentence:

"ALPGM performs different types of cross-validations to choose from: the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC) and a classical cross-validation with iterative fitting along a regularization path **(used in the case study)**."

Page 8, Line 6: "... (2) the optimizer: the method for..."; change to "...(2) the optimizer, which is the method for..."

The sentence has been rephrased as suggested by the reviewer.

Page 8, Line 6: "...(3) its (possibly nonlinear) activation functions..."; When are activation functions linear?

They are almost never linear, but it is still a possibility. For specific cases where one does not want to restrict the output values within a certain range, using a linear activation function allows to produce real values. Nonetheless, using them in more than one layer in a deep ANN does not make any sense.

The sentence has been adapted as follows:

"(3) its (**usually** nonlinear) activation functions"

Page 8, Line 10: "...allowing to train deep neural networks..."; change to "allowing the training of deep neural networks..."

The sentence has been rephrased as suggested by the reviewer.

Page 8, Line 11: "...ANNs are best suited when the quality of predictions prevails over the interpretability of the model." This is vague, and does not help readers know when ANNs are 'best suited'. How are either of these things quantified?

This cannot be strictly quantified, it depends on each field and situation. It is based on the understanding of the process which is being modelled. If a certain process is well understood, and the variables which are involved are well known, then it is acceptable to focus on prediction rather than causality. One can build a prototype model with the previous knowledge of which variables are meaningful.

The sentence has been rephrased as follows in order to clarify the sentence with respect to the goal of this study:

"As their learnt parameters are difficult to interpret, **ANN are adequate tools when the quality of predictions prevails over the interpretability of the model (the latter likely involving causal inference, sensitivity testing or modelling of ancillary variables). This is precisely**

the case in our study context here, where abundant knowledge about glacier physics further helps choosing adequate variables as input to deep learning"

Page 10, Line 21: "...should be long enough to be representative of the glacier evolution..."; How long is 'representative'? Representative of what? How does one know this?

This sentence is directly related to what has been stated previously in the same paragraph. The time difference between the two DEMs depends on the achievable signal-to-noise ratio, meaning that if a glacier is losing mass at a high pace, one will be able to use a shorter time period between the two DEMs. This is of course done with the hypothesis of glacier shrinkage in the future due to climate change, so in order to have a representative parameterization of how the glacier retreats, we need to find a period of glacier retreat in the recent past.

Due to the confusion produced by this sentence, and the fact that the necessary information is already conveyed in the same paragraph, we removed this sentence:

"As discussed in Vincent et al. (2014), the time period between the two DEMs used to calibrate the method needs to be long enough to show important ice thickness differences. The criteria will of course depend on each glacier and each period, but it will always be related to the achievable signal-to-noise ratio. Vincent et al. (2014) concluded that for their study on the Mer de Glace glacier (28.8 km2, mean altitude = 2868 m.a.s.l.) in the French Alps, the 2003-2008 period was too short, due to the delayed response of glacier geometry to a change in surface mass balance. Indeed, the results for that 5-year period diverged from the results from longer periods~~**. Moreover, the period should be long enough to be representative of the glacier evolution, which will often encompass periods with strong ablation and others with no retreat or even with positive SMBs.**~~"

Page 11, Line 3: Refer to Figure 4 here

A reference to Figure 4 has been added.

Page 11, Line 5: Is there a reference to this study?

This study is the main author's (Jordi Bolibar) PhD project, and since my PhD thesis manuscript is still to be written there is not an available reference yet.

Page 11, Line 9: "...using remote sensing based on changes in glacier volume and the snow line altitude is used..."; Remove second "is used"

The "is used" part should not be removed, otherwise the sentence would be left with a subject without verb.

"SUBJECT (An annual glacier-wide SMB dataset reconstructed using remote sensing based on changes in glacier volume and the snow line altitude) + VERB (is used)"

Commas have been added to give pause and increase clarity:

"**An annual glacier-wide SMB dataset, reconstructed using remote sensing based on changes in glacier volume and the snow line altitude, is used**"

Page 12, Figure 4: Axes should be labelled

In order to keep Fig.4 more compact in a single column, the coordinates of the bottom left corner and the top right corner have been added in the legend to guide the reader. We believe this information should be enough to properly read the map.

Page 13, Line 2: Cite RGI (check here for reference: https://www.glims.org/RGI/)

The RGI Consortium (2017) reference has been added as follows:

RGI Consortium (2017): Randolph Glacier Inventory(RGI) – A Dataset of Global Glacier Outlines: Version 6.0. Technical Report, Global Land Ice Measurements from Space, Boulder, Colorado, USA. Digital Media. Doi: 10.7265/N5-RGI-60

Page 13, Line 12: Qualifications here are vague (e.g. "quite satisfactorily", "good overall", "certain altitudinal ranges"). Give quantitative measures of "goodness", and refer to specific parts of Figure S2 that demonstrate what you're talking about.

The paragraph has been rephrased to improve the precision of the statements with references to Fig. S2:

"The simulated ice thicknesses for Saint-Sorlin (2 km^2, mean altitude = 2920 m.a.s.l., Écrins cluster) and Mer de Glace (28 km^2, mean altitude = 2890 m.a.s.l., Mont-Blanc cluster) glaciers are satisfactorily modelled by F19. **Mer de Glace's tongue presents local errors of about 50 m, peaking at 100 m (30% error) around 2000-2100 m.a.s.l, but the overall distribution of the ice is well represented. Saint Sorlin glacier follows a similar pattern, with maximum errors of around 20 m (20% error) at 2900 m.a.s.l. and a good representation of the ice distribution.**"

Page 13, Line 26: This sentence can be improved by maintaining consistency across clause structure. You use "we verb" statements (e.g. we go through, we assess, and we show) for all clauses except for "the building of the machine learning SMB models".

The sentence has been rephrased in order to keep it more consistent:

"In this section, we go through the selection of SMB predictors, **we introduce the procedure for building machine learning SMB models**, we assess their performance in space and time and we show some results of simulations using the French alpine glaciers dataset."

Page 14, Line 25 (and paragraph): You discuss that you dynamically calculate the accumulation/ablation periods based on the CPDD, and that you keep constant periods to account for winter and summer snowfalls. Later, you use 'transition months' as predictors – are these predictors kept constant, or dynamically calculated? Are results improved when the transition months are dynamically computed? I ask because I would expect that what constitutes a 'transition month' may change in the future. Or do you think that this approach, applied to more variables, then forces the model to depend too much on CPDD when the CPDD is not the only variable involved in melt?

This dynamical separation between ablation and accumulation periods is done to compute the seasonal meteorological data: the CPDD (temperature in ablation season), the winter snowfall and the summer snowfall. These three variables are introduced as climate predictors in Eq. 3. However, there are as well the monthly temperature and snowfall values in Eq. 3, which in Sect. "3.2.2 Causal analysis" are sometimes referred as transition months. The machine learning models receive all the monthly data as part of Eq. 4 and then determine which months are more relevant to explain the glacier-wide SMB of glaciers in this region. The fact that some transition months (between the ablation and accumulation periods) showed up as relevant predictors in the causal analysis, is purely based on the relationships found in the meteorological data between 1959 and 2015 for some glaciers, and between 1984 and 2015 for most glaciers of the dataset. As explained in Sect. "1 Introduction", lines 20-21, parameterized and statistical models work with the hypothesis that the relationships found in data remain stationary in time. This is of course not totally true in our case, which is why we decided to dynamically compute the ablation season (CPDD) to account for the (likely) longer ablation periods in the future. Therefore, the seasonal meteorological data adapts to future climate changes, but the individual relationships found in monthly data remain constant. Nonetheless, since there are many predictors for monthly data, their importance is very distributed, so these stationary relationships based on past climate data should not have such an important effect.

Page 15, Equations 2 and 3: Are input variables normalized? If so, how?

The input variables are only normalized for the Lasso. The ANN includes batch normalization internally, so raw data is fed directly to the input layer. This is already mentioned in "2.2.4 Deep artificial neural network", but it was indeed not specified for the Lasso. Therefore, a new line has been added in "2.2.3 Lasso" as follows:

"All input data is normalized by removing the mean and scaling to unit variance."

Page 15, Line 20: When you say 'linear machine learning', are you referring to the linear regression methods? Be consistent in how you refer to your methods.

With "linear machine learning" we mean the linear methods used in this paper (OLS and Lasso). "Linear regression methods" and "linear machine learning" are equivalent terms in this paper, since we are only working with regression. The term "linear regression" is only used as "multiple linear regression" when referring to OLS or stepwise multiple linear regression. The terms "linear machine learning" and "nonlinear machine learning" are the ones used throughout the paper, especially in Sect. 3 and 4 to refer to the differences found between linear methods and nonlinear deep learning.

In order to avoid confusion, the sentence has been changed as follows using the plural to refer to both linear machine learning models:

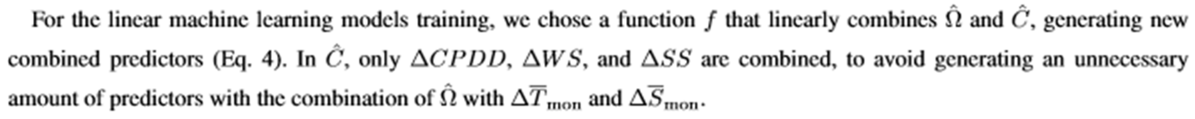
"For the linear machine learning model**s** training, we chose a function f that …"

Page 15, Line 20: How did you choose the function f?

Function f is based on the data expansion mentioned in Page 6, line 14. The idea is to linearly combine topographical and seasonal climatic data, with the exception of the monthly data. Monthly data is not combined to avoid the generation of an unnecessary number of predictors. The sentence has been adapted as follows for clarity:

"For the linear machine learning models training, we chose a function \$f\$ that linearly combines \$\hat{\Omega}\$ and \$\hat{C}\$, generating new combined predictors (Eq. 4**). In \$\hat{C}\$, only \${\Delta CPDD}\$, \${\Delta WS}\$, and \${\Delta SS}\$ are combined, to avoid generating an unnecessary amount of predictors with the combination of \$\hat{\Omega}\$ with \${\Delta \overline{T}_{\operatorname{mon}}}\$ and \${\Delta \overline{S}_{\operatorname{mon}}}\$.**"

LaTex print:



For the linear machine learning models training, we chose a function f that linearly combines $\hat{\Omega}$ and $\hat{C}$, generating new combined predictors (Eq. 4). In $\hat{C}$, only $\Delta CPDD$, ΔWS, and ΔSS are combined, to avoid generating an unnecessary amount of predictors with the combination of $\hat{\Omega}$ with $\Delta \overline{T}_{\mathrm{mon}}$ and $\Delta \overline{S}_{\mathrm{mon}}$.

Page 15, Equation 25: You create linear models using the predictors shown here. You then create nonlinear models using only the predictors in Equations 2 and 3. Then, you compare the results of these models and conclude that the nonlinear model is better because of the nonlinear nature of the model; however, how do you know that the improved performance is not simply due to using a different set of predictor variables? Your argument would be more convincing if you first showed that the linear model performance improved when you change predictor variables from the standard case (those only in Equations 2 and 3) to the combination case (Equation 4), and then showed that a nonlinear model using variables from the standard case outperformed even this improved linear model.

For both OLS and Lasso, a subset selection or coefficient shrinkage is done in order to reduce the number of kept predictors. Even in the expanded Eq. 4, the original predictors from Eq. 2 and 3 are still there, so they are potential candidates to be chosen. We believe that linear models trained with Eq. 4 can be compared to a deep ANN trained with Eq. 2 + 3, because as mentioned in Page 16, line 2, the ANN already performs all the possible combinations in each layer by combining all the input predictors. For OLS, it would not change anything to test only using Eq. 2 + 3, since we are computing all the possible combinations of Eq. 4 already, so the Eq. 2 + 3 subset is already covered in the current case. For Lasso, the situation would be slightly similar. Some early tests were already done without combination of climate and topographical predictors with worse results. Moreover, as Fig. 5 shows, the top predictor and many of the top predictors in Lasso are combined, so the added value is already shown in the results.

In order to clarify these aspects, a sentence has been added in Sect. 3.2.1 as follows:

“Early Lasso tests (not shown here) using only the predictors from Eq. 2 and 3 demonstrated the benefits of expanding the number of predictors, as it is later shown in Fig. 5.”

Page 15, Equation 5: Is there a missing ‘(‘? This equation ends with ‘)_g,y’

Indeed, there is a format problem with the parenthesis. Eq. 4. This has been fixed as suggested by the reviewer.

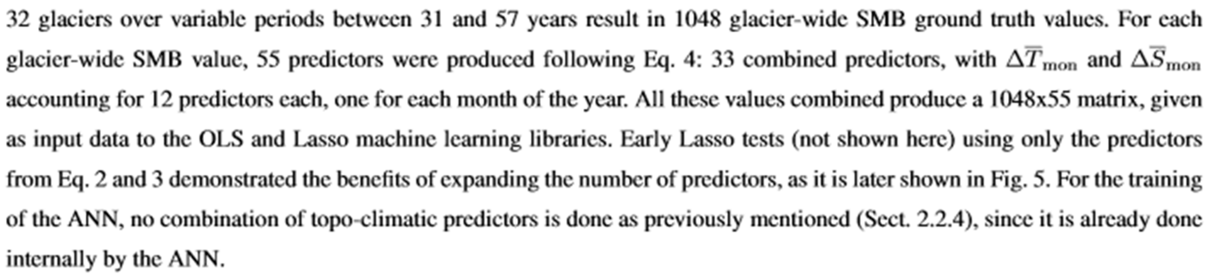
Page 16, Line 1: It is not clear to me why there are 50 predictors, when there are 33 coefficients in Equation 4.

In fact, this is a mistake. There used to be 50 predictors, but after a small change in the code they should have been updated to 55. From the 33 predictors, two are the mean monthly temperature and the monthly snowfall, which account for 12 predictors each (one value per month). Therefore: 33 – 2 + 24 = 55 predictors.

This number has been fixed throughout the manuscript and a sentence has been added to clarify this aspect:

“32 glaciers over variable periods between 31 and 57 years result in 1048 glacier-wide SMB ground truth values. **For each glacier-wide SMB value, 55 predictors were produced following Eq. 4: 33 combined predictors, with ${\Delta \overline{T}_{\operatorname{mon}}}$ and ${\Delta \overline{S}_{\operatorname{mon}}}$ accounting for 12 predictors each, one for each month of the year.** All these values combined produce a 1048x55 matrix, given as input data to the OLS and Lasso machine learning libraries.”

LaTex print:



32 glaciers over variable periods between 31 and 57 years result in 1048 glacier-wide SMB ground truth values. For each glacier-wide SMB value, 55 predictors were produced following Eq. 4: 33 combined predictors, with $\Delta\overline{T}_{\mathrm{mon}}$ and $\Delta\overline{S}_{\mathrm{mon}}$ accounting for 12 predictors each, one for each month of the year. All these values combined produce a 1048x55 matrix, given as input data to the OLS and Lasso machine learning libraries. Early Lasso tests (not shown here) using only the predictors from Eq. 2 and 3 demonstrated the benefits of expanding the number of predictors, as it is later shown in Fig. 5. For the training of the ANN, no combination of topo-climatic predictors is done as previously mentioned (Sect. 2.2.4), since it is already done internally by the ANN.

Page 16, Line 2: Can you please be more explicit about what this matrix is, and to what matrix equation it is input into?

Machine learning libraries work with matrices formed by a series of lines, one for each ground truth value used as a reference, and columns formed by the respective predictors for each ground truth value. Therefore, data generated following Eq. 4 is structured as a matrix with the respective glacier-wide SMB values, forming a 1048x55 matrix. For each of the 1048 glacier-wide SMB values, we generate 55 predictors following Eq. 4.

In order to make this point clear, the whole paragraph has been rephrased as follows:

"32 glaciers over variable periods between 31 and 57 years result in 1048 glacier-wide SMB ground truth values. For each glacier-wide SMB value, 55 predictors were produced following Eq. 4: 33 combined predictors, with ${\Delta \overline{T}_{\operatorname{mon}}}$ and ${\Delta \overline{S}_{\operatorname{mon}}}$ accounting for 12 predictors each, one for each month of the year. All these values combined produce a 1048x55 matrix, given as input data to the OLS and Lasso machine learning libraries."

LaTex print: see previous image.

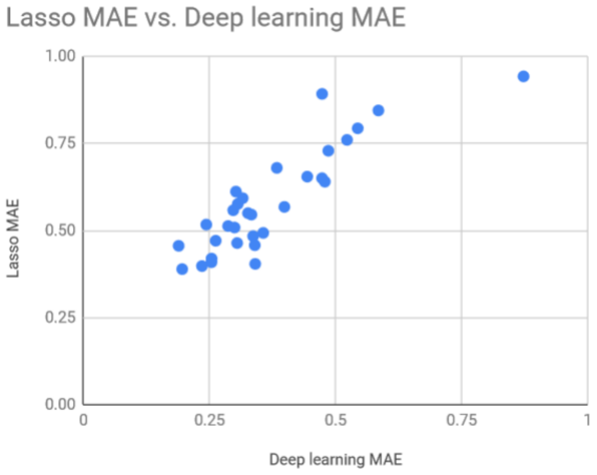
Page 16, Line 8: "...the annual CPDD as well as the winter and summer snowfall appear as significant predictors as well as several monthly mean temperatures and snowfall values..."; Change to "...the annual CPDD, winter and summer snowfall, and several monthly mean temperature and snowfall were found to be significant at p<?..."

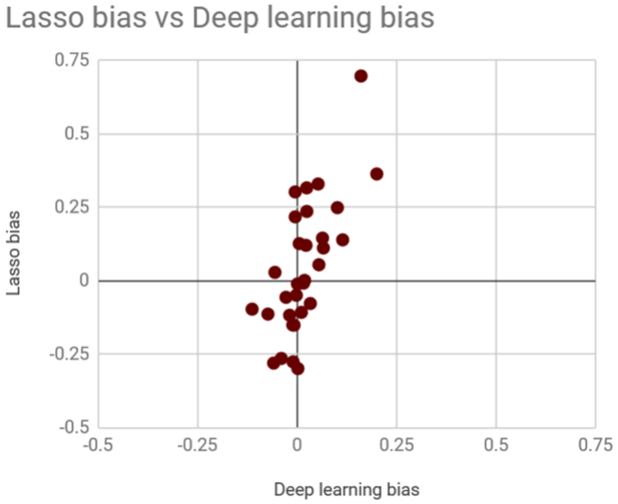
In this sentence we did not imply to use the word "significant" in statistical terms. A more accurate word choice, in the context of the causal analysis which determines the importance (%) of each predictor would be:

"Regarding the climatic variables, the annual CPDD as well as the winter and summer snowfall appear as **the most important** predictors together with several monthly mean temperatures and snowfall values (Fig. 5)"

Page 20, Figure 8: This is a challenging plot to interpret (in my opinion). Can you plot deep learning bias vs lasso bias and deep learning MAE vs Lasso MAE? That may more clearly demonstrate the points you make, and may reveal structure. The points in the scatter plots could be coloured by region (Ecrins, Vanoise, Mont Blanc) if there are regional patterns. If this approach is not useful or helpful, then the current figure will suffice.

We agree that there is a lot of information in the plot, but after testing different types of plots, we believe this is the most effective way to convey the information. As requested, we plotted both the MAE and bias between Deep learning and Lasso, but in our opinion, these plots are not clear and do not serve to establish a clear comparison between the two methods.





The confusion might come from the fact that since glaciers are structured in massifs, the reader is tempted to look for regional patterns. Since there are no clear patterns, and the plot does not intend to show that, this now has specifically been mentioned in the legend as follows:

"Figure 8. Mean average error (MAE) and bias (vertical bars) for each glacier of the training dataset structured by clusters for the 1984-2014 LOGO glacier-wide SMB simulation. **No clear regional error patterns arise**"

Page 20, Line 15: "This implies, that. . . "; remove the comma

The sentence has been adapted as suggested by the reviewer.

Page 22, Line 7: "...using Leave-Some-Glaciers-and-Years-Out (LSYGO)"; the abbreviation should be LSGYO, or the full phrase should be Leave-Some-Years-and-Glaciers-Out, for consistency

Indeed, the definition has been changed to "Leave-Some-Years-and-Glaciers-Out (LSYGO)".

Page 25, Lines 4-5: "The greater the dropout, the more we will constrain the learning of the ANN so the higher the generalization will be, until a certain point." This sentence is not clear. What does it mean to "constrain the learning"? Why is there a "certain point", and what happens beyond that point? This could be made more explicit.

As explained in "2.2.4 Deep artificial neural network", dropout is a regularization technique which consists of disconnecting certain connections in the neural network in order to reduce or constrain the amount of learning. This has been shown to help the ANN generalize. There is a range of dropout values which will produce this effect, but when pushed too far, the ANN will become too small, therefore unable to find meaningful patterns in data and performance will start to drop. The key aspect in the use of dropout is finding the good range of dropout values to make the ANN generalize without dropping performance. As with any hyperparameter tuning in ANNs, this is done via cross-validation as explained in the results section.

In order to make this point clearer, the sentence has been adjusted as follows:

"The greater the dropout, the more we will constrain the learning of the ANN so the higher the generalization will be, until a certain point**, where relevant information will start to be lost and performance will drop.**"

Page 25, Line 7: Why is it that slower convergence leads to better generalization? Is this always true?

For a given gradient descent optimizer, a slower convergence, meaning a smaller learning rate and a greater number of epochs, generally results in a better generalization. As explained in page 25, lines 6-7, a slower convergence has higher chances to encounter the global minima, whereas faster learning rates results in bigger jumps throughout the error landscape, thus often getting stuck in local minima or in certain regions of the error landscape. As for any hyperparameter, there is a range of values for which this is applicable. Depending on each case and dataset, a slower convergence might not improve generalization nor performance. Moreover, using a too slow converge will likely hamper or totally prevent any learning.

The sentence has been adjusted:

"On the other hand, the learning rate to compute the stochastic gradient descent, which tries to minimize the loss function, also plays an important role: smaller learning rates **generally** result in a slower convergence towards the absolute minima, thus producing models with better generalization."

Page 25, Line 8: ". . . that best suits a certain dataset and model." How does one define "best"?

The best model is determined using cross-validation and looking at different metrics, such as the RMSE or the coefficient of determination. The sentence has been updated to take this into account:

"By balancing all these different effects, one can achieve the accuracy versus generalization ratio that best suits a certain dataset and model **in terms of performance**"

Page 25, Line 17: "Despite it has been shown. . ."; Change to "Although it has been shown" or "Despite the fact that it has been shown"

The sentence has been modified to "Despite the fact that it has been shown…"

Page 25, Line 28: "The results were quite astonishing. . ."; If the results are astonishing, then this result warrants further emphasis in the paper. The methods used to come to this conclusion should be brought up in Section 3, and further discussion is warranted in Section 4. It is worth a figure to communicate these results

The adjective "astonishing" is probably not suitable in this context, as it might lead to exaggeration. These results are quite straightforward in terms of interpretation, since the methods are exactly the same as for the case study. The only difference is the number of input variables used. We believe that giving the numeric metrics is enough to convey the message, and such results do not deserve specific plots in an already quite long manuscript and supplementary material.

The adjective "astonishing" has been changed to "**interesting**" to reflect this.

Page 26, Lines 5-16: This paragraph is speculative, but is presented with a high degree of confidence. Phrases such as "unprecedented efficiency" and "excellent" are used without supporting evidence. Much of the discussion is implied; for example: "An interesting way of expanding a dataset would be to use a deep learning approach to fill the data gaps." It is unclear how this would be done. "Such an approach would be an excellent way of obtaining more SMB data in remote glacierized regions such as the Andes or the Himalayas." This is not known or demonstrated by the rest of the paper. I would recommend either removing this paragraph entirely or severely limiting its scope.

Indeed, this paragraph contains a lot of propositions, some not directly related to the results found in this paper. Some of the sentences have been removed to be more straight to the point and to avoid any speculation. Nonetheless, the first sentence "An interesting way of expanding a dataset would be to use a deep learning approach to fill the data gaps", is the direct consequence of all the methodology presented here. The relationships found and learnt in data can then be extrapolated to other glaciers and periods to make predictions. That is the main reason we have done such a thorough cross-validation in both the spatial and temporal dimensions. We have no studies to use as a reference for this, since our results of this regional SMB reconstruction will soon be submitted as a separate paper, and apparently there are no other similar studies yet which have used such an approach. We believe this claims are valid, since the performance and viability of this approach is precisely proven in this methodological paper, for which we show the performance of this method compared to other classical approaches (multiple linear regression). Indeed, in other regions with smaller data coverage it might differ, but being in a discussion section, we think it is important to state the potential of this approach in these regions. Even if has been only tested in the European Alps for now, it would be extremely interesting to do so in regions such as the Andes or the High Mountains of Asia.

In order to deal with all these statements raised here, the paragraph has been widely updated:

"Deep learning can be of special interest once applied in the reconstruction of SMB time series. More and more SMB data is becoming available thanks to the advances in remote sensing sing (e.g., Brun et al. (2017); Rabatel et al. (2017); Zemp et al. (2019)), but these datasets often cover limited areas and the most recent time period in the studied regions. An interesting way of expanding a dataset would be to use a deep learning approach to fill the data gaps, **based on the relationships found in a subset of glaciers as in the case study presented here. Past SMB time series of vast glacierized regions could thereby be reconstructed, with potential applications in remote glacierized regions such as the Andes or the Himalayas**. ~~**It could also be applied in data-rich regions benefiting from regionalized climate reanalyses (e.g. Caillouet et al. (2016)), covering the 1871-present period for France). Another possibility would be to completely bypass both the SMB and glacier dynamics of a classic glacier evolution model by training a deep ANN which would directly simulate changes in glacier thicknesses. If the ANN is trained with enough glacier thickness changes, climatic and topographical data, it could be able to simulate the 3D evolution of the glacier straight from the raw data. It might still be too soon for such models to be implemented, but once enough data will be available in the future, this could be a promising new way of tackling glacier evolution modelling.**~~"

Page 26, Lines 18-22: This paragraph is speculative. It does not follow from the results presented in the paper, and is more of a justification for using deep learning in glaciology than it is an item of discussion in the context of the preceding research. These final two paragraphs do a disservice to the rest of the paper; prior to this, the organization had nice flow, and the first two paragraphs in Section 4.3 were both interesting and directly relevant.

We agree that this last paragraph is not directly related to the work presented here, and was included in the broader context of deep learning applied in glaciology. In order to keep the pace of the discussion and to simplify the discussion this whole last paragraph has been deleted from the manuscript.

Page 27, Lines 4-12: This paragraph is quite vague and does not explicitly follow from the research. For example: "It might still be too early for the development of such models in certain regions" is a vague statement. Conclusions should follow directly and explicitly from the work and should not reach beyond the scope of the research. The first paragraph in Section 5 is much better.

This paragraph focuses on the applications of the methodology presented here. As mentioned in a previous comment, these applications are just a consequence of the work and results presented in the French Alps case study. ALPGM, as a model and as a tool is capable of reconstructing and simulating glacier-wide SMBs at regional scale. We believe it is important to state why a model such ALPGM can be useful and what are its potential applications. These applications and their results will be presented as two separate papers, which show the results of applying deep learning for glacier-wide SMB reconstruction and for future glacier evolution predictions, using it as a SMB model (alternative to temperature index).

In order to improve the fluidity, and to relate all the statements to the research presented in this paper, the paragraph has been rephrased as follows:

"Deep learning should be seen as an opportunity by the glaciology community. Its good performance **for SMB modelling** in both the spatial and temporal dimensions shows how relevant it can be for a broad range of applications. Combined with in situ or remote sensing SMB estimations, it can serve to reconstruct SMB time series for regions or glaciers with already available data for past and future periods, **with potential applications in remote regions such as the Andes or the high mountains of Asia**. Moreover, deep learning can be used as an alternative to classical SMB models **as it is done in ALPGM**: important nonlinearities from the glacier and climate systems are potentially ignored by these mostly linear models, which could give an advantage to deep learning models in regional studies. It might still be too early for the development of such models in certain regions **which lack consistent datasets with a good spatial and temporal coverage**. **Nevertheless**, as new data becomes available the gap is slowly being closed towards real big data approaches in glaciology."

Supplementary Figures: In the SMB_lasso_ANN_no_weights_SMB_simulations.pdf file, y-axes are missing units.

The supplementary figures' y-axes have been updated as suggested by the reviewer.

Figures 6, 7, 9, and 11: Please increase font size, especially of axis labels.

The font size has been increased as suggested by the reviewer.

Deep learning applied to glacier evolution modelling

Jordi Bolibar[1,2], Antoine Rabatel[1], Isabelle Gouttevin[3], Clovis Galiez[4], Thomas Condom[1], and Eric Sauquet[2]

[1]Univ. Grenoble Alpes, CNRS, IRD, G-INP, Institut des Géosciences de l'Environnement (IGE, UMR 5001), Grenoble, France
[2]Irstea, UR RiverLy, Lyon-Villeurbanne, France
[3]Univ. Grenoble Alpes, Université de Toulouse, Météo-France, CNRS, CNRM, Centre d'Études de la Neige, Grenoble, France
[4]Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, Grenoble, France

Correspondence: Jordi Bolibar (jordi.bolibar@univ-grenoble-alpes.fr)

Abstract. We present a ~~parameterized glacier evolution model, with a~~ novel approach to simulate and reconstruct annual glacier-wide surface mass balance (SMB) ~~component~~ series based on a deep artificial neural network (*i.e.* deep learning). This method has been included as the SMB component of an open-source regional glacier evolution model. While most glacier models tend to incorporate more and more physical processes, here we take an alternative approach by creating a parameterized model based on data science. Annual glacier-wide SMBs can be simulated from topo-climatic predictors using either deep learning or Lasso (regularized multilinear regression), whereas the glacier geometry is updated using a glacier-specific parameterization. We compare and cross-validate our nonlinear deep learning SMB model against other standard linear statistical methods on a dataset of 32 French alpine glaciers. Deep learning is found to outperform linear methods, with improved explained variance (up to +64% in space and +108% in time) and accuracy (up to +47% in space and +58% in time), resulting in an estimated r^2 of 0.77 and RMSE of 0.51 m.w.e. Substantial nonlinear structures are captured by deep learning, with around 35% of nonlinear behaviour in the temporal dimension. For the glacier geometry evolution, the main uncertainties come from the ice thickness data used to initialize the model. These results should encourage the use of deep learning in glacier modelling as a powerful nonlinear tool, capable of capturing the nonlinearities of the climate and glacier systems, that can serve to reconstruct or simulate SMB time series for individual glaciers ~~at regional scale~~ in a whole region for past and future climates.

1 Introduction

Glaciers are arguably one of the most important icons of climate change, being climate proxies which can ~~clearly~~ depict the evolution of climate for the global audience ~~(IPCC (2018)). For~~ (IPCC, 2018). In the coming decades, mountain glaciers will be some of the most important contributors to sea level rise and will most likely drive important changes in the hydrological regime of glaciarized catchments ~~(Beniston et al. (2018); Vuille et al. (2018); Hock et al. (2019))~~(Beniston et al., 2018; Vuille et al., 2018; Hock e . The reduction in ice volume may produce an array of ~~consequences~~ hydrological, ecological and economic consequences in mountain regions which requires to be properly predicted. These consequences will strongly depend on the future climatic scenarios, which will determine the timing and magnitude for the transition of hydrological regimes ~~(Huss and Hock (2018)).~~ (Huss and Hock, 2018). Understanding these future transitions is key for societies to adapt to future hydrological and climate configurations.

Glacier and hydro-glaciological models can help answer these questions, giving several possible outcomes depending on multiple climate scenarios. ~~For any glacier model, two main processes have to be simulated: the glacier surface~~ (a) Surface mass balance (SMB) and ~~the glacier dynamics ; which can be done using approaches of different complexity~~ (b) glacier dynamics both need to be modelled to understand glacier evolution on regional and sub-regional scales. Models of varying complexity exist for both processes. In order to model these processes at large scale (*i.e.* ~~.~~on several glaciers at a catchment scale), some compromises need to be made, which can be approached in different ways:

(a) Regarding SMB:

1. Empirical models, like the temperature-index model (e.g. Hock, 2003), simulate glacier SMB through empirical relationships between air temperature and melt and snow accumulation.

2. Statistical or machine learning models describe and predict glacier SMB based on statistical relationships found in data from a selection of topographical and climate predictors (e.g. Martin, 1974; Steiner et al., 2005).

3. Physical and Surface Energy Balance (SEB) models take into account all energy exchanges between the glacier and the atmosphere, and can simulate the spatial and temporal variability of snowmelt and the changes in albedo (e.g. Gerbaux et al., 2005) .

(b) Regarding glacier dynamics:

1. Parameterized models do not explicitly resolve any physical processes, but implicitly take them into account using parameterizations, based on statistical or empirical relationships, in order to modify the glacier geometry. This type of models range from very simple statistical models ~~(Carlson et al. (2014))~~ (e.g. Carlson et al., 2014) to more complex ones based on different approaches, such as a calibrated equilibrium-line altitude (ELA) model ~~(Zemp et al. (2006))~~ (e.g. Zemp et al., 2006), a glacier retreat parameterization specific for glacier size groups (Huss and Hock, 2015) or volume/length-area scaling ~~(Marzeion et al. (2012); Radić et al. (2014))~~(e.g. Marzeion et al., 2012; Radić et al., 2014).

2. Process-based models, like ~~GloGEM (Huss and Hock (2015)), GloGEMflow (Zekollari et al. (2019)) and OGGM (Maussion et al. (2~~ ~~)~~GloGEMflow (e.g. Zekollari et al., 2019) and OGGM (e.g. Maussion et al., 2019), approximate a number of glacier physical processes ~~and surface energy and mass balance in a simplified manner~~involved in ice flow dynamics using the shallow ice approximation.

3. Physics-based models, like the finite elements Elmer/Ice model ~~Gagliardini et al. (2013)~~(e.g. Gagliardini et al., 2013), approach glacier dynamics by explicitly simulating physical processes and solving the full Stokes equations ~~Jouvet et al. (2009); Réve~~ ~~. They are generally not suitable for regional studies due to their high complexity.~~ (e.g. Jouvet et al., 2009; Réveillet et al., 2015) .

At the same time, the use of these different approaches strongly depend on available data, whose spatial and temporal resolutions have an important impact on the results' quality and uncertainties ~~(*e.g.*, Réveillet et al. (2018)). Parameterized models~~ (e.g., Réveillet et al., 2018). Parameterized glacier dynamics models and empirical and statistical SMB models require a reference or training dataset to calibrate the relationships, which can then be used for projections with the hypothesis that relationships remain stationary in time. On the contrary, process-based and specially physics-based glacier dynamics and SMB models have the advantage of representing physical processes, but they require larger datasets at higher spatial and temporal resolutions with a consequently higher computational cost ~~(Réveillet et al. (2018)). Meteorological~~ (Réveillet et al., 2018). For SMB modelling, meteorological reanalyses provide an attractive alternative to sparse point observations, although their spatial resolution and suitability to complex high-mountain topography are often not good enough for high-resolution physics-based glacio-hydrological applications. However, parameterized models are much more flexible, equally dealing with fewer and coarser meteorological data as well as the state of the art reanalyses, which allows to work at resolutions much closer to glaciers' scale and to reduce uncertainties. The current resolution of climate projections is still too low to adequately drive most glacier physical processes, but the ever-growing datasets of historical data are paving the way for the training of parameterized machine learning models.

In glaciology, statistical models have been applied for more than half a century, starting with simple multiple linear regressions on few meteorological variables ~~(Hoinkes (1968); Martin (1974))~~(Hoinkes, 1968; Martin, 1974). Statistical modelling has made enormous progress in the last decades, specially thanks to the advent of machine learning. Compared to other fields in geosciences, such as oceanography (e.g., Ducournau and Fablet, 2016; Lguensat et al., 2018), climatology (e.g., Rasp et al., 2018; Jiang et and hydrology (e.g., Marçais and de Dreuzy, 2017; Shen, 2018), we believe that the glaciological community has ~~remained quite oblivious to these advances, mostly focusing on glacier physics and physical-based and process-based models~~not yet exploited the full capabilities of these approaches. Despite this fact, a number of studies have taken steps towards statistical approaches. ~~Steiner et al. (2005)~~ Steiner et al. (2005) pioneered the very first study to use artificial neural networks (ANNs) in glaciology to simulate mass balances of the Grosse Aletschgletscher in Switzerland. They showed that a nonlinear model is capable of better simulating glacier mass balances compared to a conventional stepwise multiple linear regression. Furthermore, they found a significant nonlinear part within the climate/glacier mass balance relationship. This work was continued in ~~Steiner et al. (2008) and Nussbaumer et al. (2012)~~ Steiner et al. (2008) and Nussbaumer et al. (2012) for the simulation of glacier length instead of mass balances. Later on, ~~Maussion et al. (2015)~~ Maussion et al. (2015) developed an empirical statistical downscaling tool based on machine learning in order to retrieve glacier surface energy and mass balance (SEB/SMB) fluxes from large-scale atmospheric data. They used different machine learning algorithms, but all of them were linear, which are not necessarily the most suitable for modelling the nonlinear climate system ~~(Houghton et al. (2001))~~(Houghton et al., 2001). Nonetheless, more recent developments in the field of machine learning and optimization enabled the use of deeper network structures than the 3-layer ANN of ~~Steiner et al. (2005)~~Steiner et al. (2005). These deeper ANNs, which remain unexploited in glaciology, allow to capture more nonlinear structures in the data even for relatively small datasets ~~(Ingrassia and Morlini (2005); Olson et al. (2018))~~(Ingrassia and Morlini, 2005; Olson et al., 2018).

Here, we present a parameterized regional open-source glacier model: the ALpine Parameterized Glacier Model ~~(ALPGM, Bolibar (2019))~~(ALPGM, Bolibar, 2019). When most glacier evolution models tend to incorporate more and more physical processes ~~(Maussion et al. (2019); Zekollari et al. (2019))~~in SMB or ice dynamics (e.g., Maussion et al., 2019; Zekollari et al., 2019) , ALPGM takes an alternative approach based on data science for SMB modelling and parameterizations for glacier dynamics simulation. ALPGM simulates annual glacier-wide SMB and the evolution of glacier volume and surface area over time scales from a few years to a century at a regional scale. Glacier-wide SMBs are computed using a deep ANN, fed by several topographical and climatic variables, an approach which is compared to different linear methods in the present paper. In order to distribute these annual glacier-wide SMBs and to update the glacier geometry, a refined version of the Δh methodology ~~(*e.g.*, Huss et al. (2008))~~ (e.g., Huss et al., 2008) is used, for which we dynamically compute glacier-specific Δh functions. In order to validate this approach, we use a case study with 32 French alpine glaciers for which glacier-wide annual SMBs are available over the period 1984-2014 and 1959-2015 for certain glaciers. High resolution meteorological reanalyses for the same time period are used ~~(SAFRAN: Durand et al. (2009))~~ (SAFRAN, Durand et al., 2009) while the initial ice thickness distribution of glaciers are taken from ~~Farinotti et al. (2019)~~Farinotti et al. (2019), for which we performed a sensitivity analysis based on field observations.

In the next section, we present an overview of the proposed glacier evolution model framework with a detailed description of the two components used to simulate the annual glacier-wide SMB and the glacier geometry update. Then, a case study using French alpine glaciers is presented, which enables to illustrate an example of application of the proposed framework including a rich dataset, the parameterized functions, as well as the results and their performance. In the end, several aspects regarding machine and deep learning modelling in glaciology are discussed, from which we make some recommendations and draw the final conclusions.

2 Model overview and methods

In this section we present an overview of the ALPGM glacier model. Moreover, the two components of this model are presented in detail: the Glacier-wide SMB Simulation component and the Glacier Geometry Update component.

2.1 Model overview and workflow

ALPGM is an open-source glacier model coded in Python. The source code of the model is accessible in the project repository (see Code availability). It is structured in multiple files which execute specific separate tasks. The model can be divided into two main components: (1) the Glacier-wide SMB Simulation and (2) the Glacier Geometry Update. The Glacier-wide SMB Simulation component is based on machine learning, taking both meteorological and topographical variables as inputs. The Glacier Geometry Update component generates the glacier-specific parameterized functions and modifies annually the geometry of the glacier (*e.g.* ice thickness distribution, glacier outline) based on the glacier-wide SMB ~~values simulated~~ models generated by the Glacier-wide SMB simulation component.

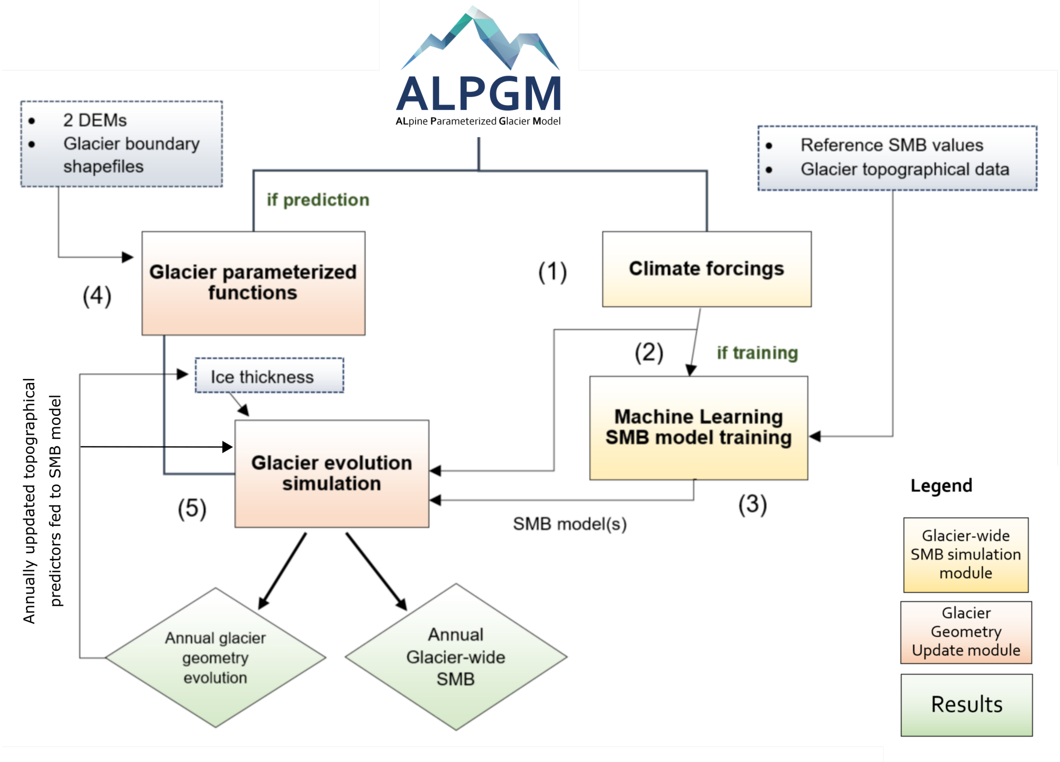

ALPGM
ALpine Parameterized Glacier Model
2 DEMs
Glacier boundary shapefiles
Reference SMB values
Glacier topographical data
if prediction
Glacier parameterized functions
(1)
Climate forcings
(4)
(2)
if training
Ice thickness
Machine Learning SMB model training
Annually uppdated topographical predictors fed to SMB model
Glacier evolution simulation
(5)
(3)
SMB model(s)
Legend
Glacier-wide SMB simulation module
Glacier Geometry Update module
Results
Annual glacier geometry evolution
Annual Glacier-wide SMB

Figure 1. ALPGM structure and workflow

Fig. 1 presents ALPGM's basic workflow. The workflow execution can be configured via the model interface, allowing to run or skip any of the following steps:

1. The meteorological forcings are preprocessed in order to extract the necessary data closest to each glacier's centroid. The meteorological features are stored in intermediate files in order to reduce computation times for future runs, automatically skipping this preprocessing step when the files have already been generated.

2. The SMB machine learning component retrieves the preprocessed ~~meteorological features and assembles the spatiotemporal training dataset , comprised by both climatic and topographical data. An~~ climate predictors from the stored files, retrieves the topographical predictors from the multitemporal glacier inventories, and then it assembles the training dataset by combining all the necessary topo-climatic predictors. A machine learning algorithm is chosen for the SMB model, which can be loaded from a previous run or it can be trained again with a new dataset. Then, the SMB model(s) are trained with the full topo-climatic dataset. These model(s) are stored in intermediate files, allowing to skip this step for future runs.

3. Performances of the SMB models can be evaluated with a leave-one-glacier-out (LOGO) or a leave-one-year-out (LOYO) cross-validation. This step can be skipped when using already established models. Basic statistical performance metrics

are given for each glacier and model, as well as plots with the simulated cumulative glacier-wide SMBs compared to their reference values with uncertainties for each of the glaciers from the training dataset.

4. The Glacier Geometry Update component starts with the generation of the glacier specific parameterized functions, using a raster containing the difference of the two pre-selected digital elevation models (DEMs) covering the study area for two separate dates, as well as the glacier contours. These parameterized functions are then stored in individual files to be used in the final simulations.

5. Once all previous steps have been run, the ~~final~~ glacier evolution simulations are launched. For each glacier, the initial ice thickness ~~raster and the parameterized~~ and DEM rasters and the glacier geometry update function are retrieved. ~~The meteorological data at the glaciers' centroid is re-computed with an annual time step based on each~~ Then, in a loop, for every glacier and year, the topographical data is computed from these raster files. The climate predictors at the glacier's ~~evolving topographical characteristics. These forcings are used to simulate the annual~~ current centroid are retrieved from the climate data (e.g. reanalysis or projections) and with all this data the input topo-climatic data for the glacier-wide SMB ~~using the machine learning model . Once an annual~~ model is assembled. Afterwards, the glacier-wide SMB ~~value is obtained, the changes in geometry are computed using the parameterized function , thus~~ for this glacier and year is simulated, which combined with the glacier-specific geometry update function allows to update the glacier's ice thickness and DEM rasters. This process is repeated in a loop, therefore updating the glacier's ~~DEM and ice thickness rasters~~geometry with an annual timestep and taking into account the glacier's morphological and topographical changes in the glacier-wide SMB simulations. For the simulation of the following year's SMB, the previously updated ice thickness and DEM rasters is used to re-compute the topographical parameters, which in turn are used as input topographical predictors for the glacier-wide SMB machine learning model. If all the ice thickness raster pixels of a glacier become zero, the glacier is considered as disappeared and is removed from the simulation pipeline. For each year, multiple results are stored in data files as well as the raster DEM and ice thickness values for each glacier.

2.2 Glacier-wide surface mass balance simulation

Annual glacier-wide SMBs are simulated using machine learning. ~~Although the features used as input for the model are classical descriptors of the topographical and meteorological conditions of the glaciers, it isworth mentioning that applying the model in different areas or with different data sources would likely require a re-training of the model due to possible biases: different regions on the globe may have other descriptors of importance but also different measuring techniques will likely have different biases~~Due to the regional characteristics and specificities of topographical and climate data, this glacier-wide SMB modelling method is, for now, a regional approach.

2.2.1 Selection of explanatory topographical and climatic variables

In order to narrow down which topographical and climatic variables best explain glacier-wide SMB in a given study area, a literature review as well as a statistical sensitivity analysis are performed. Typically used topographical predictors are longitude,

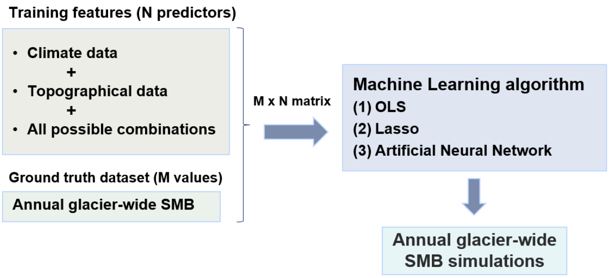

Training features (N predictors)
• Climate data
+
• Topographical data
+
• All possible combinations
Ground truth dataset (M values)
Annual glacier-wide SMB
M x N matrix
Machine Learning algorithm
(1) OLS
(2) Lasso
(3) Artificial Neural Network
Annual glacier-wide SMB simulations

Figure 2. Glacier-wide SMB simulation component workflow. Machine learning models are dynamically created based on training data

latitude, glacier slope and mean altitude. As for meteorological predictors, cumulative positive degree days (CPDD), but also mean monthly temperature, snowfall and possibly other variables that influence the surface energy budget are often used in the literature. Examples of both topographic and meteorological predictors can be found in the case study in Sect. 3. A way to prevent biases when making predictions with different climate data is to work with anomalies, calculated as differences of the variable with respect to its average value over a chosen reference period.

For the machine learning training, the relevant predictors must be selected, ~~performing~~ so we perform a sensitivity study of the annual glacier-wide SMB to topographical and climatic variables over the study training period. This can be performed with individual linear regressions between each variable and glacier-wide SMB data. After identification of the topographical and climatic variables that can potentially explain annual glacier-wide SMB variability for the region of interest, a training dataset is built. An effective way of expanding the training dataset in order to dig deeper into the available data is to combine the climatic and topographical input variables (Weisberg, 2014). Such combinations can be expressed following Eq. (1):

$$SMB_{g,y} = f(\hat{\Omega},\hat{C}) + \varepsilon_{g,y} \tag{1}$$

Where $\hat{\Omega}$ is a vector of the selected topographical predictors, $\hat{C}$ is a vector with the selected climatic features and $\varepsilon_{g,y}$ is the residual error for each annual glacier-wide SMB value, $SMB_{g,y}$.

Once the training dataset is created, different algorithms f (two linear and one nonlinear, for the case of this study) can be chosen to create the SMB model: (1) OLS (Ordinary Least Squares) all-possible multiple linear regressions; (2) Lasso (Least absolute shrinkage and selection operator) ~~(Tibshirani (1996))~~(Tibshirani, 1996); and (3) a deep Artificial Neural Network (ANN). ALPGM uses some of the most popular machine learning Python libraries: ~~StatsModel (Seabold and Perktold (2010))~~StatsModels (Seabold and Perktold, 2010), Scikit-learn ~~(Pedregosa et al. (2012)) and Keras (Chollet (2015))~~ (Pedregosa et al., 2012) and Keras (Chollet, 2015) with a TensorFlow backend. The overall workflow of the machine learning glacier-wide SMB model production in ALPGM is summarized in Fig. 2.

2.2.2 All-possible multiple linear regressions

With the ordinary least squares (OLS) all-possible multiple linear regressions, we attempt to find the best subset of predictors in Eq. 1 based on the resulting r^2 adjusted, while at the same time avoiding overfitting ~~(Hawkins (2004))~~ (Hawkins, 2004) and collinearity, and limiting the complexity of the model. As its name indicates, the goal is to minimize the residual sum of squares for each subset of predictors ~~(Hastie et al. (2009))~~(Hastie et al., 2009). n models are produced by selecting all possible subsets of k predictors. It is advisable to narrow down the number of predictors for each subset in the search to reduce the computational cost. Models with low performance are filtered out, keeping only models with highest r^2 adjusted possible, a variance inflation factor (VIF) < 1.2 and a p-value < 0.01/n (in order to ensure the Bonferroni correction). Retained models are combined by averaging their predictions, thereby avoiding the pitfalls related to stepwise single model selection ~~(Whittingham et al. (2006))~~(Whittingham et al., 2006). These criteria ensure that the models explain as much variability as possible, avoid collinearity and are statistically significant.

2.2.3 Lasso

The Lasso (Least absolute shrinkage and selection operator) ~~(Tibshirani (1996))~~ (Tibshirani, 1996) is a shrinkage method which attempts to overcome the shortcomings of the simpler step-wise and all-possible regressions. In these two classical approaches, predictors are discarded in a discrete way, giving subsets of variables which have the lowest prediction error. However, due to its discrete selection, these different subsets can exhibit high variance, which does not reduce the prediction error of the full model. The Lasso performs a more continuous regularization by shrinking some coefficients and setting others to zero, thus producing more interpretable models ~~(Hastie et al. (2009))~~(Hastie et al., 2009). Because of its properties, it strikes a balance between subset selection (like all-possible regressions) and Ridge regression ~~(Hoerl and Kennard (1970)).~~ (Hoerl and Kennard, 1970) . All input data is normalized by removing the mean and scaling to unit variance. In order to determine the degree of regularisation applied to the coefficients used in the linear OLS regression, an alpha parameter needs to be chosen using cross-validation. ALPGM performs different types of cross-validations to choose from: the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC) and a classical cross-validation with iterative fitting along a regularization path (used in the case study). Alternatively, a Lasso model with Least Angle Regression, also known as Lasso Lars ~~(Tibshirani et al. (2004))~~(Tibshirani et al., 2004), can also be chosen with a classical cross-validation. ~~All the input data is scaled and centered to zero before training the model. The generated coefficients from the model serve to determine the significant predictors to be kept for the artificial neural network training.~~

2.2.4 Deep artificial neural network

Artificial neural networks (ANNs) are nonlinear statistical models inspired by biological neural networks ~~(Fausett (1994); Hastie et al. (2009~~ ~~)~~(Fausett, 1994; Hastie et al., 2009). A neural network is characterized by: (1) the architecture or pattern of connections between units and the number of layers (input, output and hidden layers); (2) the optimizer: which is the method for determining the weights of the connections between units; and (3) its (~~possibly~~ usually nonlinear) activation functions ~~(Fausett (1994)~~

~~)~~(Fausett, 1994). When ANNs have more than one hidden layer (*e.g.* Fig. 3), they are referred to as deep ANNs or deep learning. The description of neural networks is beyond the scope of this study, so for more details and a full explanation please refer to ~~Fausett (1994), Hastie et al. (2009)~~Fausett (1994), Hastie et al. (2009), as well as ~~Steiner et al. (2005, 2008)~~ Steiner et al. (2005, 2008) where the reader can find a thorough introduction to the use of ANNs in glaciology. ANNs gained recent interest thanks to improvements of optimization algorithms allowing ~~to train~~ the training deep neural networks, that lead to better representation of complex data patterns. As their learnt parameters are difficult to interpret, ~~ANNs are best suited~~ ANN are adequate tools when the quality of predictions prevails over the interpretability of the model ~~.~~ (the latter likely involving causal inference, sensitivity testing or modelling of ancillary variables). This is precisely the case in our study context here, where abundant knowledge about glacier physics further helps choosing adequate variables as input to deep learning. Their ability to model complex functions of the input parameters makes them particularly suitable for modelling complex nonlinear systems such as the climate system ~~(Houghton et al. (2001))~~ (Houghton et al., 2001) and glacier systems ~~(Steiner et al. (2005))~~(Steiner et al., 2005).

ALPGM uses a feedforward fully-connected ANN (Fig. 3). In such an architecture, the processing units - or neurons - are grouped into layers where all the units of a given layer are fully connected to all units of the next layer. The flow of information is directional, from the input layer (*i.e.*. in which each neuron corresponds to one of the N explanatory variables) to the output neuron (*i.e.*. corresponding to the target variable of the model, the SMB). For each connection of the ANN, weights are initialized in a random fashion following a specific distribution (generally centred around 0). In each unit of each hidden layer, the weighted values are summed before going through a nonlinear activation function, responsible for introducing the nonlinearities in the model. Using a series of iterations known as epochs, the ANN will try to minimize a specific loss function (the mean squared error (MSE) in our case) comparing the processed values of the output layer with the ground truth (y). In order to avoid falling into local minima of the loss function, some regularisation is needed to prevent the ANN from overfitting ~~(Hastie et al. (2009))~~(Hastie et al., 2009). To prevent overfitting during the training process (*i.e.*. to increase the ability of the model to generalize to new data), we used a classical regularization method called dropout, consisting in training iteratively smaller subparts of the ANN by randomly disconnecting a certain amount of connections between units. The introduction of Gaussian noise at the input of the ANN also helped to generalize, as it performs a similar effect to data augmentation. The main consequence of regularisation is generalization, for which the produced model is capable of better adapting to different configurations of the input data.

The hyperparameters used to configure the ANN are determined using cross-validation, in order to find the best performing combination of number of units, hidden layers, activation function, learning rate and regularisation method. Due to the relatively small size of our dataset, we encountered the best performances with a quite small deep ANN, with a total of 6 layers (4 hidden layers) with a (N, 40, 20, 10, 5, 1) architecture (Fig. 3), where N is the number of selected features. Since the ANN already performs all the possible combinations between features (predictors), we use a reduced version of the training matrix from Eq. 1, with no combination of climatic and topographical features. Due to the relatively small size of the architecture, the best dropout rates are small ~~(Srivastava et al. (2014))~~(Srivastava et al., 2014), and range between 0.3 and 0.01 depending on the number of units of each hidden layer. Leaky ReLUs have been chosen as the activation function, because of their widespread

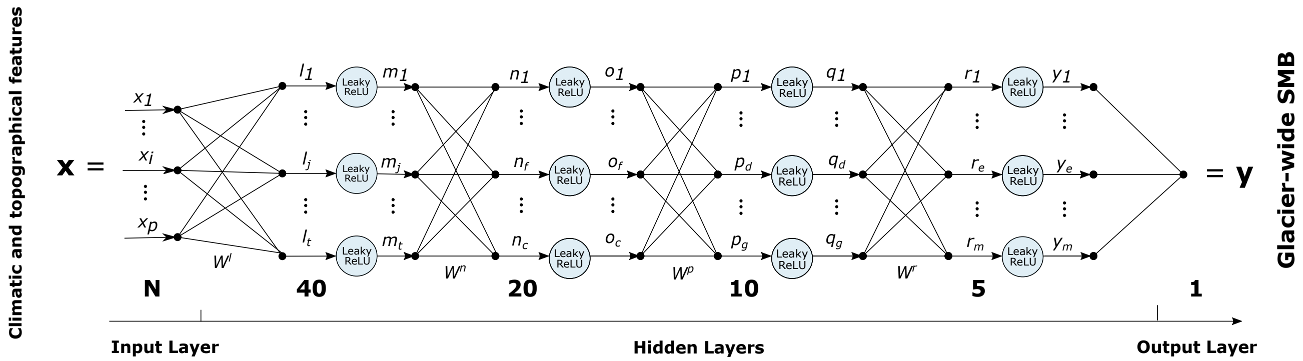


Figure 3. Deep Artificial Neural Network architecture used in ALPGM. The numbers indicate the number of neurons in each layer

reliability and the fact they help prevent the "dead ReLU" problem, where certain neurons can stop "learning" ~~(Xu et al. (2015))~~(Xu et al., 2015). The He uniform initialization ~~(He et al. (2015))~~ (He et al., 2015) has been used as it is shown to work well with Leaky ReLUs, and all unit bias were initialized to zero. In order to optimize the weights of the gradient descent, we used the RMSprop optimizer, for which we fine-tuned the learning rate, obtaining the best results at 0.0005 in space and 0.02 in time. Each batch was normalized before applying the activation function in order to accelerate the training ~~(Ioffe and Szegedy (2015))~~(Ioffe and Szegedy, 2015).

Like for many other geophysical processes found in nature, extreme annual glacier-wide SMB values occur much less often than average values, approximately following an unbounded Gumbel-type distribution ~~(Thibert et al. (2018))~~(Thibert et al., 2018) . From a statistical point of view, this means that ANN will "see" few extreme values and will accord less importance to them. For future projections in a warmer climate, extreme positive glacier-wide SMB balances should not be the main concern of glacier models. However, extreme negative annual glacier-wide SMB values should likely increase in frequency, so it is in the modeller's interest to reproduce them as well as possible. Setting the sample weights as the inverse of the probability density function during the ANN training can partly compensate for the imbalance of a dataset. This boosts the performance of the model for the extreme values, at the cost of sacrificing some performance on more average values, which can be seen as a r^2/RMSE trade-off (see Fig. 6 and 9 from the case study). The correct setting of the sample weights allows the modeller to adapt the ANN to each dataset and application.

2.3 Glacier geometry update

Since the first component of ALPGM simulates annual glacier-wide SMBs, these changes in mass need to be redistributed over the glacier surface-area in order to reproduce glacier dynamics. This redistribution is applied using the Δh parameterization. The idea was first developed by ~~Jóhannesson et al. (1989)~~ Jóhannesson et al. (1989) and then adapted and implemented by ~~Huss et al. (2008)~~Huss et al. (2008). The main idea behind it is to use two or more DEMs covering the study area. These DEMs should have dates covering a period long enough (which will be later discussed in detail). By subtracting them, the changes in glacier surface elevation over time can be computed, which corresponds to a change in thickness (considering no basal erosion). Then, these thickness changes are normalized and considered as a function of the normalized glacier altitude. This Δh function is specific for each glacier and represents the normalized glacier thickness evolution over its altitudinal range. One advantage of such a parametrized approach is that it implicitly considers the ice flow which redistributes the mass from the accumulation to the ablation area. In order to make the glacier volume evolve in a mass-conserving fashion, we apply this function to the annual glacier-wide SMB values in order to scale and distribute its change in volume.

As discussed in ~~Vincent et al. (2014)~~Vincent et al. (2014), the time period between the two DEMs used to calibrate the method needs to be long enough to show important ice thickness differences. The criteria will of course depend on each glacier and each period, but it will always be related to the achievable signal-to-noise ratio. ~~Vincent et al. (2014)~~ Vincent et al. (2014) concluded that for their study on the Mer de Glace glacier (28.8 km^2, mean altitude = 2868 m.a.s.l.) in the French Alps, the 2003-2008 period was too short, due to the delayed response of glacier geometry to a change in surface mass balance. Indeed, the results for that 5-year period diverged from the results from longer periods. Moreover, the period should be long enough to be representative of the glacier evolution, which will often encompass periods with strong ablation and others with no retreat or even with positive SMBs.

Therefore, by subtracting the two DEMs, the ice thickness difference is computed for each specific glacier. These values can then be classified by altitude, thus obtaining an average glacier thickness difference for each pixel altitude. As a change to previous studies ~~(Vincent et al. (2014); Huss and Hock (2015); Hanzer et al. (2018); Zekollari et al. (2019))~~(Vincent et al., 2014; Huss and Hoc , we no longer work with altitudinal transects, but with individual pixels. In order to filter noise and artefacts coming from the DEM raster files, different filters are applied to remove outliers and pixels with unrealistic values, namely at the border of glaciers or where the surface slopes are high (refer to Supplements for detailed information). Our methodology thus allows to better exploit the available spatial information based on its quality, and not on arbitrary location within transects.

3 Case study: French alpine glaciers

3.1 Data

All data used in this case study is based on the French Alps (Fig. 4), located in the westernmost part of the European Alps, between 5.08° and 7.67°E, and 44° and 46°13’N. This region is particularly suited for the validation of a glacier evolution model because of the wealth of available data. Moreover, ALPGM has been developed as part of a hydro-glaciological study to understand the impact of the retreat of French alpine glaciers in the Rhône river catchment (97,800 km^2).

3.1.1 Glacier-wide surface mass balance

An annual glacier-wide SMB dataset, reconstructed using remote sensing based on changes in glacier volume and the snow line altitude~~is used (Rabatel et al. (2016))~~, is used (Rabatel et al., 2016). This dataset is constituted by annual glacier-wide SMB values for 30 glaciers in the French Alps (Fig. 4) for 31 years, between 1984-2014. The great variety in topographical characteristics of the glaciers included in the dataset, with a good coverage of the three main clusters or groups of glaciers in

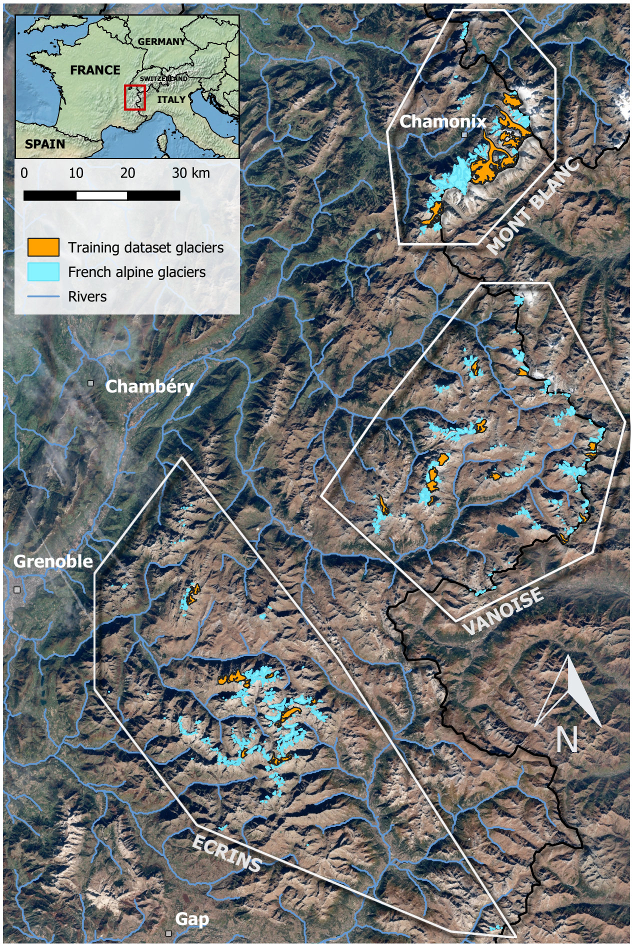


Figure 4. French alpine glaciers used for model training and validation and their classification into 3 clusters/regions (Écrins, Vanoise, Mont-Blanc). Coordinates of bottom left map corner: 44º32'N, 5º40'E, coordinates of the top right map corner: 46º08'N, 7º17'E.

the French Alps (Fig. 4), makes them an ideal training dataset for the model. Each of the clusters represents a different setup of glaciers with different contrasting latitudes (Écrins and Mont-Blanc), longitudes (Écrins and Vanoise), glacier size (smaller glaciers in Écrins and Vanoise *vs* larger ones in Mont-Blanc) and climatic characteristics with a Mediterranean influence towards the south of the study region. For more details regarding this dataset refer to ~~Rabatel et al. (2016)~~Rabatel et al. (2016).

5 Data from the Mer de Glace, Saint-Sorlin, Sarennes and Argentière glaciers is also used, coming from field observations from the GLACIOCLIM observatory. For some of these glaciers, glacier-wide SMB values are available since 1949, although only values from 1959 onwards were used to match the meteorological reanalysis. This makes a total of 32 glaciers (Argentière and Saint-Sorlin glaciers belonging to the two datasets), representing 1048 annual glacier-wide SMB values (taking into account some gaps in the dataset).

3.1.2 Topographical glacier data and altimetry

~~All topographical data for this study has been~~ The topographical data used for the training of the glacier-wide SMB machine learning models is taken from the multitemporal inventory of the French Alps glaciers ~~(*e.g.*, Gardent et al. (2014))~~ (e.g., Gardent et al., 2014 partly available through the GLIMS Glacier Database ~~(NSIDC (2005))~~(NSIDC, 2005). We worked with the 1967, 1985, 2003 and 2015 inventories ~~(Gardent et al. (2014), with 2015 update). Two~~ (Gardent et al., 2014, with 2015 update). Between these dates, the topographical predictors are linearly interpolated. On the other hand, in the glacier evolution component of ALPGM (Fig.1, step 5), the topographical data is re-computed every year for each glacier from the evolving and annually updated glacier-specific ice thickness and DEM rasters (Sect. 3.1.3). Since these raster files are estimates for the year 2003 (Farinotti et al. (2019) for the ice thickness), the full glacier evolution simulations can start the earliest at this date. For the computation of the glacier-specific geometry update functions, two DEMs covering the whole French Alps have been used: (1) one from 2011 generated from SPOT5 stereo-pair images, acquired on 15 October 2011; and (2) a 1979 aerial photogrammetric DEM from the French National Geographic Institute (Institut Géographique National, IGN), processed from aerial photographs taken around 1979. Both DEMs have an accuracy between 1 and 4 meters ~~(Rabatel et al. (2016))~~(Rabatel et al., 2016), and their uncertainties are negligible compared to many other parameters in this study.

3.1.3 Glacier ice thickness

Glacier ice thickness data come from ~~Farinotti et al. (2019)~~Farinotti et al. (2019), hereafter F19, based on the Randolph Glacier Inventory ~~(RGIv6~~v6.0 ~~)~~(RGI, Consortium, 2017). The ice thickness values represent the latest consensus estimate, averaging an ensemble of different methods based on the principles of ice flow dynamics to invert the ice thickness from surface characteristics.

We also have ice thickness data acquired by diverse field methods ~~(seismic, ground penetrating radar or hot water drilling, Rabatel et al. (2018))~~ (seismic, ground penetrating radar or hot water drilling, Rabatel et al., 2018) for four glaciers of the GLACIOCLIM observatory. We compared these in situ thickness data, with the simulated ice thicknesses from F19 (refer to Supplements for detailed information). Although differences can be found (locally up to 100% in the worst cases), no systematic biases were found with respect to glacier local slope nor glacier altitude; therefore, no systematic correction was applied to the dataset. The simulated ice thicknesses for Saint-Sorlin (2 km^2, mean altitude = 2920 m.a.s.l., Écrins cluster) and Mer de Glace (28 km^2, mean altitude = 2890 m.a.s.l., Mont-Blanc cluster) glaciers are ~~quite~~ satisfactorily modelled by F19~~, with a good overall agreement except for the local distributionin certain altitudinal ranges~~. Mer de Glace's tongue presents local errors of about 50 m, peaking at 100 m (30% error) around 2000-2100 m.a.s.l, but the overall distribution of the ice is well represented. Saint Sorlin glacier follows a similar pattern, with maximum errors of around 20 m (20% error) at 2900 m.a.s.l. and a good representation of the ice distribution. The ice thicknesses for Argentière Glacier (12.8 km^2, mean altitude = 2808 m.a.s.l., Mont-Blanc cluster) and Glacier Blanc (4.7 km^2, mean altitude = 3196 m.a.s.l., Écrins cluster) are underestimated by F19 with an almost constant bias with respect to altitude, as seen in ~~Rabatel et al. (2018)~~Rabatel et al. (2018). Therefore, a manual correction was applied to the F19 datasets for these two glaciers based on the field observations from the GLACIOCLIM observatory. A detailed plot (Fig. S2) presenting these results can be found in the supplementary material.

3.1.4 Climate data

In our French Alps case study, ALPGM is forced with daily mean near-surface (2 m) temperatures, daily cumulative snowfall and rain. The SAFRAN dataset is used to provide this data close to the glaciers' centroids. SAFRAN meteorological data ~~(Durand et al. (2009))~~ (Durand et al., 2009) is a reanalysis of weather data including observations from different networks, and specific to the French mountain regions (Alps, Pyrenees and Corsica). Instead of being structured as a grid, data is provided at the scale of massifs, which are in turn divided into altitude bands of 300 meters and into 5 different aspects (north, south, east, west and flat).

3.2 Glacier-wide surface mass balance simulations: validation and results

In this section, we go through the selection of SMB predictors, ~~the building of the~~ we introduce the procedure for building machine learning SMB models, we assess their performance in space and time and we show some results of simulations using the French alpine glaciers dataset.

3.2.1 Selection of predictors

Statistical relationships between meteorological and topographical variables with respect to glacier-wide SMB are frequent in the literature for the European Alps ~~(Hoinkes (1968)). Martin (1974)~~ (Hoinkes, 1968). Martin (1974) performed a sensitivity study on the SMB of the Saint-Sorlin and Sarennes glaciers (French Alps) with respect to multi-annual meteorological observations for the 1957-1972 period. ~~Martin (1974)~~ Martin (1974) obtained a multiple linear regression function based on annual precipitation and summer temperatures, and he concluded that it could be further improved by differentiating winter and summer precipitations. ~~Six and Vincent (2014)~~ Six and Vincent (2014) studied the sensitivity of the SMB to climate change in the French Alps from 1998 until 2014. They found that the variance of summer SMB is responsible for over 90% of the variance of the annual glacier-wide SMB. ~~Rabatel et al. (2013, 2016)~~ Rabatel et al. (2013, 2016) performed an extensive sensitivity analysis of different topographical variables (slope of the lowermost 20% of the glacier area, mean elevation, surface area, length, minimum elevation, maximum elevation, surface area change and length change) with respect to glacier ELA and annual glacier-wide SMBs of French alpine glaciers. Together with ~~Huss (2012)~~Huss (2012), who performed a similar study with SMB, the most significant statistical relationships were found for the lowermost 20% area slope, the mean elevation, glacier surface area, aspect and easting and northing. ~~Rabatel et al. (2013)~~ Rabatel et al. (2013) also determined that the climatic interannual variability is mainly responsible for driving the glacier equilibrium-line altitude temporal variability, whereas the topographical characteristics are responsible for the spatial variations in the mean ELA.

Summer ablation is often accounted for by means of cumulative positive degree days (CPDD). However, in the vast majority of studies, accumulation and ablation periods are defined between fixed dates (*e.g.*, 1st October - 30th April for the accumulation period in the northern mid-latitudes) based on optimizations. As discussed in ~~Zekollari and Huybrechts (2018)~~ Zekollari and Huybrechts (2018), these fixed periods may not be the best to describe SMB variability through statistical correlation. Moreover, the ablation season will likely evolve in the coming century, due to climate warming. In order to overcome these limitations, we dynamically calculate each year the transition between accumulation and ablation seasons (and vice-versa) based on a chosen quantile in the CPDD (Fig. S3). We found higher correlations between annual SMB and ablation-period CPDD calculated using this dynamical ablation season. On the other hand, it was not the case for the separation between summer and winter snowfall. Therefore, we decided to keep constant periods to account for winter (1st October-1st May) and summer (1st May-1st October) snowfalls, and to keep them dynamical for the CPDD calculation.

Following this literature review, vectors ~~and~~ $\hat{\Omega}$ and $\hat{C}$ from (Eq. 1) read as:

$$\hat{\Omega} = \begin{bmatrix} \overline{Z} & Z_{\max} & \alpha_{20\%} & \text{Area} & \text{Lat} & \text{Lon} & \Phi \end{bmatrix} \tag{2}$$

$$\widehat{C} = \begin{bmatrix} \Delta CPDD & \Delta WS & \Delta SS & \Delta\overline{T}_{\text{mon}} & \Delta\overline{S}_{\text{mon}} \end{bmatrix} \tag{3}$$

Where:

$\overline{Z}$: Mean glacier altitude

$Z_{\max}$: Maximum glacier altitude

$\alpha_{20\%}$: Slope of the lowermost 20% glacier altitudinal range

$Area$: Glacier surface area

Lat: Glacier latitude

Lon: Glacier longitude

Φ: Cosine of the glacier's aspect (North = 0º)

$\Delta CPDD$: CPDD (Cumulative Positive Degree Days) anomaly

ΔWS: Winter snow anomaly

ΔSS: Summer snow anomaly

$\Delta\overline{T}_{\text{mon}}$: Average temperature anomaly for each month for the hydrological year

$\Delta\overline{S}_{\text{mon}}$: Average snowfall anomaly for each month for the hydrological year

For the linear machine learning ~~model~~ models training, we chose a function f that linearly combines $\hat{\Omega}$ and $\hat{C}$, generating new combined predictors (Eq. 4)~~:~~. In $\hat{C}$, only $\Delta CPDD$, ΔWS, and ΔSS are combined, to avoid generating an unnecessary amount of predictors with the combination of $\hat{\Omega}$ with $\Delta\overline{T}_{\mathrm{mon}}$ and $\Delta\overline{S}_{\mathrm{mon}}$.

$$
\begin{aligned}
SMB_{g,y} = ((&a_1\overline{Z} + a_2Z_{\max} + a_3\alpha_{20\%} + a_4Area + a_5Lat + a_6Lon + a_7\Phi + a_8)\Delta CPDD + \\
&(b_1\overline{Z} + b_2Z_{\max} + b_3\alpha_{20\%} + b_4Area + b_5Lat + b_6Lon + b_7\Phi + b_8)\Delta SS + \\
&(c_1\overline{Z} + c_2Z_{\max} + c_3\alpha_{20\%} + c_4Area + c_5Lat + c_6Lon + c_7\Phi + c_8)\Delta WS + \\
&d_1\overline{Z} + d_2Z_{\max} + d_3\alpha_{20\%} + d_4Area + d_5Lat + d_6Lon + d_7\Phi + d_8 + d_n\Delta\overline{T}_{\mathrm{mon}} + d_m\Delta\overline{S}_{\mathrm{mon}} + \varepsilon)_{\mathrm{g,y}}
\end{aligned}
\tag{4}
$$

~~This combination produces a total of 50 predictors for our~~ 32 glaciers over variable periods between 31 and 57 years ~~, resulting in a 1048x50 matrixused for~~ result in 1048 glacier-wide SMB ground truth values. For each glacier-wide SMB value, 55 predictors were produced following Eq. 4: 33 combined predictors, with $\Delta\overline{T}_{\mathrm{mon}}$ and $\Delta\overline{S}_{\mathrm{mon}}$ accounting for 12 predictors each, one for each month of the year. All these values combined produce a 1048x55 matrix, given as input data to the OLS and Lasso ~~.~~ machine learning libraries. Early Lasso tests (not shown here) using only the predictors from Eq. 2 and 3 demonstrated the benefits of expanding the number of predictors, as it is later shown in Fig. 5. For the training of the ANN, no combination of topo-climatic ~~features~~ predictors is done as previously mentioned ~~.~~ (Sect. 2.2.4), since it is already done internally by the ANN.

3.2.2 Causal analysis

By running the Lasso algorithm on the dataset based on Eq. ~~4~~2 and 3, we obtain the contribution of each predictor in order to explain the annual glacier-wide SMB variance. ~~Interestingly, not all topographical variables found in the literature were kept by Lasso, which only found significant linear relationships with the average slope of the lowermost 20% of the glacier and with the latitude and longitude.~~ Regarding the climatic variables, ~~the annual CPDD~~ accumulation-related predictors (winter snowfall, summer snowfall as well as ~~the winterand summer snowfall appear as significant predictorsas well as several monthly mean temperatures and snowfall values (Fig. 5). Latitude and longitude seem to play an important role when combined with snowfall. Indeed, glaciers located in the eastern part of the Vanoise cluster~~ several winter, spring and even summer months), appear as the most important predictors. Ablation-related predictors also seem to be relevant, mainly with CPDD and summer and shoulder season months (Fig. ~~4)are likely more affected by eastern precipitation fluxes, whereas the western glaciers in the Écrins and Mont-Blanc clusters are mostly affected by western disturbances. Latitude plays an important role as well, with Mont-Blanc glaciers receiving higher amounts of snowfall compared to the Mediterranean-influenced ones in Écrins. These findings are in agreement with Rabatel et al. (2013) and the general snow climatology in the French Alps (Durand et al. (2009)). On the other hand, the meteorological conditions are also~~ 5). Interestingly, meteorological conditions in the transition months are crucial for the annual glacier-wide SMB in the French Alps~~, especially in the transition months:~~ : (1) October temperature is determinant for the transition between the ablation and the accumulation season, favouring a lengthening of melting when temperature remains positive, or conversely allowing snowfalls that protect the ice and contribute to the accumulation when temperatures are negative; (2) March snowfall has a similar effect: positive anomalies contribute to the total accumulation at the glacier surface, and a thicker snow pack will delay the snow/ice transition during the ablation season leading to a less negative ablation rate ~~(e. g., Fig. 6b in Réveillet et al. (2018)).~~ (e.g. Fig. 6b, Réveillet et al., 2018). Therefore, meteorological conditions of these transition months seem to strongly impact the annual glacier-wide SMB variability, since their variability oscillates between positive and negative values, unlike the months in the heart of summer or winter.

In a second term, topographical predictors do play a role, albeit a secondary one. The slope of the 20% lowermost altitudinal range, the glacier area, the glacier mean altitude and aspect help to modulate the glacier-wide SMB signal, which unlike point or altitude-dependent SMB, partially depends on glacier topography (Huss et al., 2012). Moreover, latitude and longitude are among the most relevant topographical predictors, which for this case study are likely to be used as bias correctors of precipitation of the SAFRAN climate reanalysis. SAFRAN is suspected of having a precipitation bias, with higher uncertainties for high altitude precipitations (Vionnet et al., 2016). Since the French Alps present an altitudinal gradient, with higher altitudes towards the eastern and the northern massifs, we found that the coefficients linked to latitude and longitude enhanced glacier-wide SMBs with a north-east gradient.

3.2.3 Spatial predictive analysis

In order to evaluate the performance of the machine learning SMB models in space, we perform a leave-one-glacier-out (LOGO) cross-validation. For relatively small datasets like the one used in this study, cross-validation ensures that the model is validated on the full dataset. Such validation aims at understanding the model's performance for predictions on other glaciers for the same time period as during the training.

An important aspect is the comparison between linear and nonlinear machine learning algorithms used in this study. ~~Steiner et al. (2005)~~ Steiner et al. (2005) already proved that a nonlinear ANN improved the results with respect a classic step-wise multiple linear regression. Here, we draw a similar comparison using more advanced methods for a larger dataset: OLS and Lasso as linear machine learning algorithms and a deep ANN as a nonlinear one. We observed significant differences between OLS, Lasso and deep learning, both in terms of explained variance (r^2) and accuracy (RMSE) of predicted glacier-wide SMBs. On average, we found improvements between +55% and +61% in the explained variance (from 0.49 to 0.76-0.79) using the nonlinear deep ANN compared to Lasso, whereas the accuracy was improved up to 45% (from 0.74 to 0.51-0.62). This means that 27% more variance is explained with a nonlinear model in the spatial dimension for glacier-wide SMB in this region. See Fig. 6 for a full summary of the results.

An interesting consequence of the nonlinearity of the ANN is the fact that it better captures extreme SMB values compared to a linear model. A linear model can correctly approximate the main cluster of values around the median, but the linear approximation performs poorly for extreme annual glacier-wide SMB values. The ANN solves this problem, with an increased explained variance which translates into a better accuracy for extreme SMB values, even without the use of sample weights (Fig. 6).

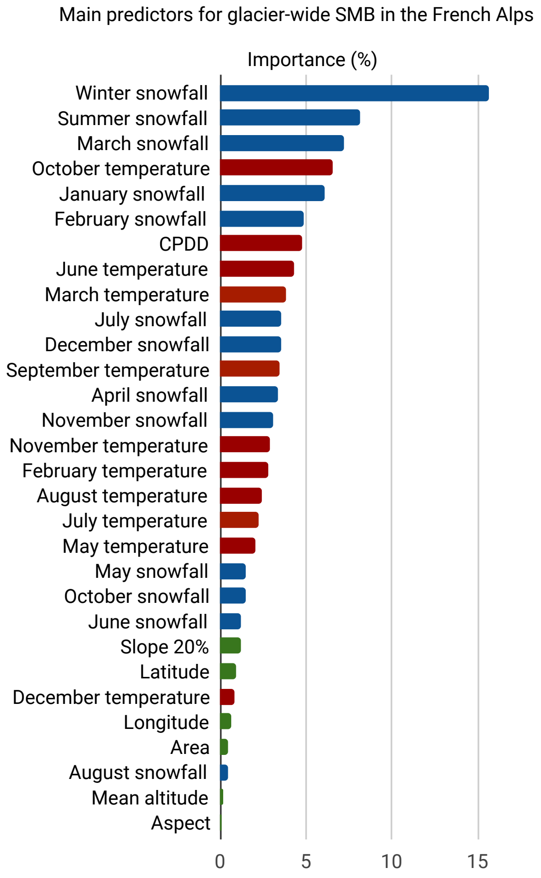


Figure 5. Contribution to the total variance of the ~~25~~ 30 top topo-climatic predictors out of ~~50~~ 55 predictors using Lasso. Green bars indicate predictors including topographical features, blue ones including accumulation-related features, and red ones including ablation-related features

As a consequence, the added value of deep learning is especially relevant on glaciers with steeper annual changes in glacier-wide SMB (Fig. 7a). The use of sample weights can scale up or down this factor, thus playing with a performance trade-off depending on how much one wants to improve the model's behaviour for extreme SMB values.

Overall, deep learning results in a lower error throughout all the glaciers in the dataset when evaluated using LOGO cross-validation (Fig. 8). Moreover, the bias is also systematically reduced, but it is strongly correlated to the one from Lasso.

3.2.4 Temporal predictive analysis

In order to evaluate the performance of the machine learning SMB models in time, we perform a leave-one-year-out (LOYO) cross-validation. This validation serves to understand the model's performance for past or future periods outside the training

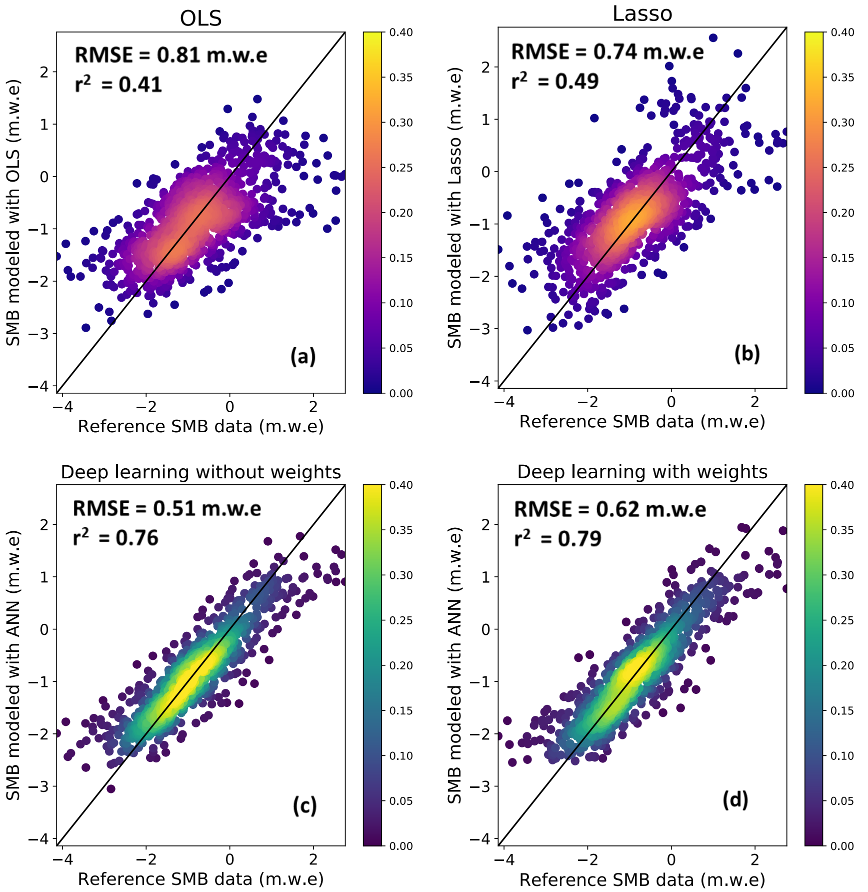


Figure 6. Evaluation of modelled annual glacier-wide SMB against the ground truth SMB data (both in m.w.e. ~~$a-1$~~a^{-1}) using Leave-One-Glacier-Out cross-validation. The colour (purple-orange for linear; blue-green for nonlinear) indicates frequency based on the probability density function. The black line indicates the reference one-to-one line. a) Scatter plot of the OLS model results; b) Scatter plot of the Lasso linear model results; Scatter plots of the deep artificial neural network nonlinear models without (c) and with sample weights (d)

time period. The best results achieved for Lasso make no use of any monthly average temperature or snowfall, suggesting that these features are not relevant for temporal predictions unlike the spatial case.

As in Sect. 3.2.3, the results between the linear and nonlinear machine learning algorithms were compared. Interestingly, using LOYO, the differences between the different models were even greater than for spatial validation, revealing the more complex nature of the information in the temporal dimension. As illustrated by Fig. 9, we found remarkable improvements between the linear Lasso and the nonlinear deep learning in both the explained variance (between +94% and +108%) and accuracy (between +32% and +58%). This implies ~~,~~that 35% more variance is explained using a nonlinear model in the temporal dimension for glacier-wide SMB balance in this region. Deep learning manages to keep very similar performances

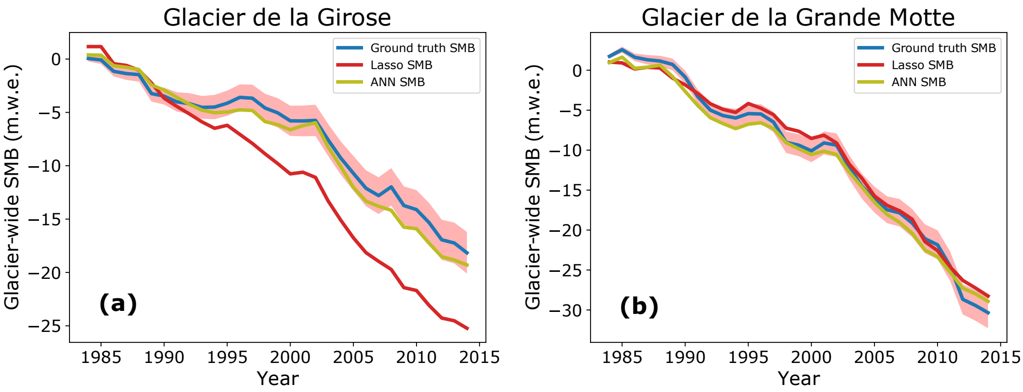


Figure 7. Examples of cumulative glacier-wide SMB (m.w.e.) simulations against the ground truth SMB data. The pink envelope indicates the accumulated uncertainties from the ground truth data. The deep learning SMB model has not been trained with sample weights in these illustrations.

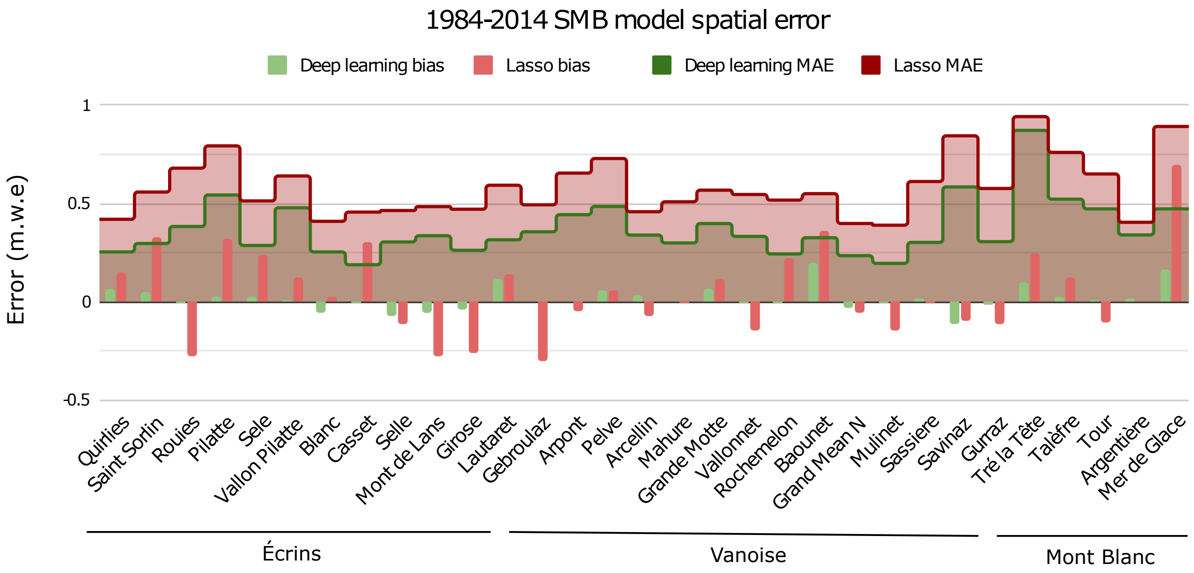


Figure 8. Mean average error (MAE) and bias (vertical bars) for each glacier of the training dataset structured by clusters for the 1984-2014 LOGO glacier-wide SMB simulation. No clear regional error patterns arise

between the spatial and temporal dimensions, whereas the linear methods see their performance affected most likely due to the increased nonlinearity of the SMB reaction to meteorological conditions.

A more detailed year by year analysis reveals interesting information about the glacier-wide SMB data structure. As seen in Fig. 10, the years with the worst deep learning precision are 1984, 1985 and 1990. All these three hydrological years present

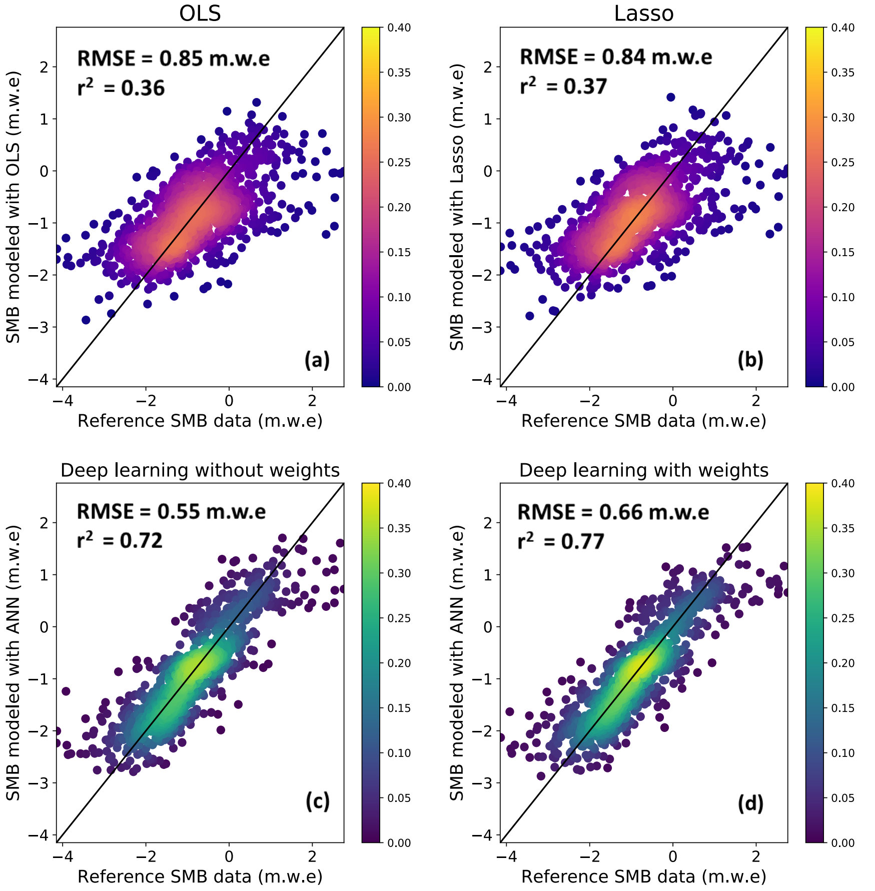


Figure 9. Evaluation of modelled annual glacier-wide SMB against the ground truth SMB data (both in m.w.e~~a^{-1}~~. a^{-1}) using Leave-One-Year-Out cross-validation. The colour (purple-orange for linear; blue-green for nonlinear) indicates frequency based on the probability density function. The black line indicates the reference one-to-one line. a) Scatter plot of the OLS model results; b) Scatter plot of the Lasso linear model results; Scatter plots of the deep artificial neural network nonlinear models without (c) and with sample weights (d).

a high spatial variability in observed (or remotely-sensed) SMBs: very positive SMB values in general for 1984 and 1985 with few slightly negative values, and extremely negative SMB values in general for 1990 with few almost neutral values. These complex configurations are clearly outliers within the dataset, which push the limits of the nonlinear patterns found by the ANN. The situation becomes even more evident with Lasso, which struggles to resolve these complex patterns and often performs poorly where the ANN succeeds (~~e.g.~~*e.g.*, years 1996, 2012 or 2014). The important bias present only with Lasso is representative of its lack of complexity towards nonlinear structures~~: the~~, which results in an underfitting of the data. The average error is ~~often good~~not bad, but it ~~is constantly biased unlike the ANN~~shows a high negative bias for the first half of

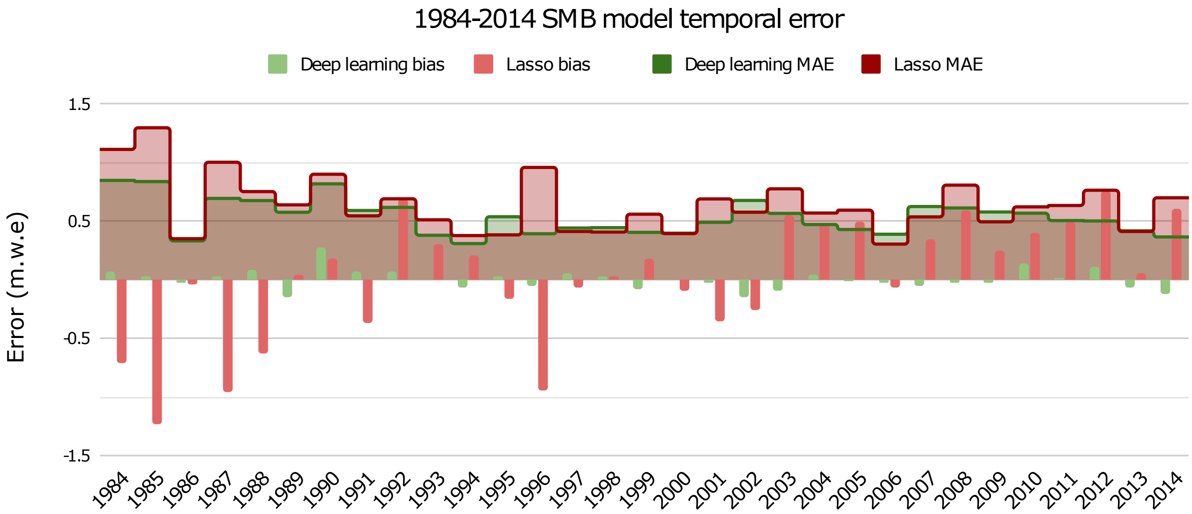


Figure 10. Mean average error (MAE) and bias (vertical bars) for each year of the training dataset for the 1984-2014 LOYO glacier-wide SMB simulation.

the period, which mostly has slightly negative glacier-wide SMBs, and a high positive bias for the second half of the period, which mostly has very negative glacier-wide SMB values.

3.2.5 Spatiotemporal predictive analysis

Once the specific performances in the spatial and temporal dimensions have been assessed, the performance in both dimensions at the same time is evaluated using ~~Leave-Some-Glaciers-and-Years-Out~~ Leave-Some-Years-and-Glaciers-Out (LSYGO) cross-validation. 64 folds were built, with test folds being comprised of data for 2 random glaciers on 2 random years, and train folds of all the data except the 2 years (for all glaciers) and the 2 glaciers (for all years) present in the test fold. These combinations are quite strict, implying that for every 4 tested values we need to drop between 123 and 126 values for training, depending on the glacier and year, to respect the spatiotemporal independence ~~(Roberts et al. (2017))~~(Roberts et al., 2017).

The performance of LSYGO is similar to LOYO, with a RMSE of 0.51 m.w.e. and a coefficient of determination of 0.77 (Fig. S5). This is reflected in the fact that very similar ANN hyperparameters were used for the training. This means that the deep learning SMB model is successful in generalizing and it does not overfit the training data.

3.3 Glacier geometry evolution: Validation and results

As mentioned in Sect. 2.3, the Δh parameterization has been widely used in many studies ~~(e. g., Huss et al. (2008, 2010); Vincent et al. (201~~ ~~).~~ (e.g., Huss et al., 2008, 2010; Vincent et al., 2014; Huss and Hock, 2015, 2018; Hanzer et al., 2018; Vincent et al., 2019). It is not in the scope of this study to evaluate the performance of this method, but we present the approach developed in ALPGM to compute the Δh functions and show some examples for single glaciers to illustrate how these glacier-specific functions perform compared to observations. For the studied French alpine glaciers, the 1979-2011 period is used. This period was proved by ~~Vincent et al. (2014)~~ Vincent et al. (2014) to be representative of Mer de Glace's secular trend. Other sub-periods could have been used, but it was shown that they did not necessarily improve the performance. In addition, the 1979 and 2011 DEMs are the only ones available that cover all the French alpine glaciers. Within this period, some years with neutral to even positive surface mass balances in the late 1970s and early 1980s can be found, as well as a remarkable change from 2003 onward with strongly negative surface mass balances, following the heatwave that severely affected the western Alps in summer 2003.

The glacier-specific Δh functions are computed for glaciers $\geq$ 0.5 km^2, which represented about 80% of the whole glaciarized surface of the French Alps in 2015 (some examples are illustrated in the Supplement Fig. S4). For the rest of very small glaciers (< 0.5 km^2), a standardized flat function is used in order to make them shrink equally at all altitudes. This is done to simulate the fact that generally, the equilibrium line of very small glaciers has surpassed the glacier's maximum altitude, thus shrinking from all directions and altitudes in summer. Moreover, due to their reduced size and altitudinal range, the ice flow no longer has the same importance as for larger or medium sized glaciers.

In order to evaluate the performance of the parameterized glacier dynamics of ALPGM, coupled with the glacier-wide SMB component, we compared the simulated glacier area of the 32 studied glaciers with the observed area in 2015 from the most up-to-date glacier inventory in the French Alps. Simulations were started in 2003, for which we used the F19 ice thickness dataset. In order to take into account the ice thickness uncertainties, we ran three simulations with different versions of the initial ice thickness: the original data, -30% and +30% of the original ice thickness in agreement with the uncertainty estimated by the authors. Moreover, in order to take into account the uncertainties in the Δh glacier geometry update function computation, we added a ±10% variation in the parameterized functions (Fig. 11).

Overall, the results illustrated in Fig. 11 show a good agreement with the observations. Even for a 12-year period, the initial ice thickness remains the largest uncertainty, with almost all glaciers falling within the observed area when taking it into account. The mean error in simulated surface area was of ~~12~~10.7% with the original F19 ice thickness dataset. Other studies using the Δh parameterization already proved that the initial ice thickness is the most important uncertainty in glacier evolution simulations, together with the choice of a GCM for future projections ~~(Huss and Hock (2015))~~(Huss and Hock, 2015).

4 Discussion and perspectives

4.1 Linear methods still matter

Despite the fact that deep learning often outperforms linear machine learning and statistical methods, there is still a place for such methods in modelling. Indeed, unlike ANNs, simpler regularised linear models such as Lasso allow an easy interpretation of the coefficients associated to each input feature, which helps to understand the contribution of each of the chosen variables to the model. This means that linear machine learning methods can be used for both prediction and causal analysis. Training a linear model in parallel to an ANN has therefore the advantage to provide a simpler linear alternative which can be used to understand the dataset. Moreover, seeing the contribution of each coefficient, one can reduce the complexity of the dataset by

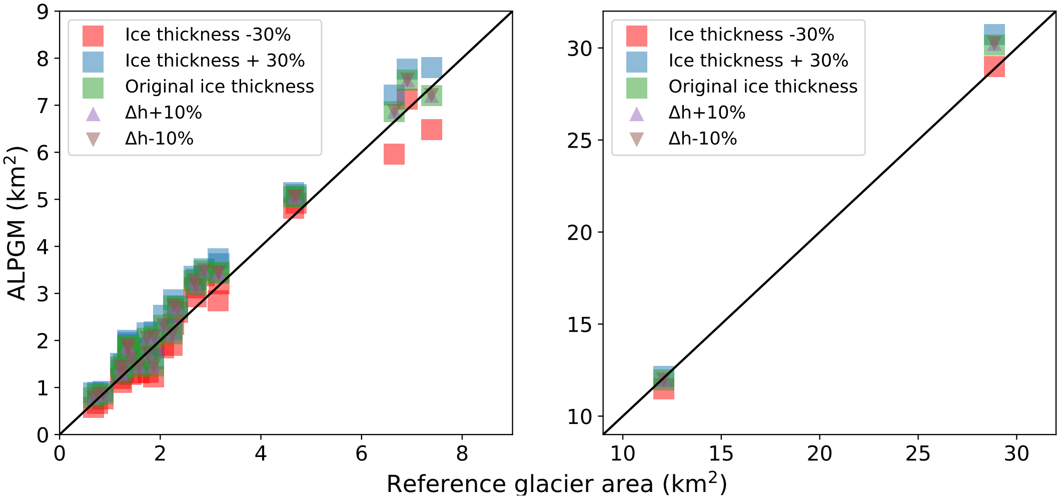


Figure 11. Simulated glacier areas for the 2003-2015 period for the 32 study glaciers using a deep learning SMB model without weights. ~~Colours~~ Squares indicate the different F19 initial ice thicknesses used taking into account their uncertainties and triangles indicate the uncertainties linked to the glacier-specific geometry update functions. For better visualisation, the figure is split in two with the two largest French glaciers on the right.

keeping only the most significant predictors. Finally, a linear model serves as well as a reference to highlight and quantify the nonlinear gains obtained by deep learning.

4.2 Training deep learning models with spatiotemporal data

The creation and training of a deep ANN requires a certain knowledge and strategy with respect to the data and study focus. When working with spatiotemporal data, the separation between training and validation becomes tricky. The spatial and temporal dimensions in the dataset cannot be ignored, and strongly affect the independence between training and validation data ~~(Roberts et al. (2017); Oliveira et al. (2019))~~(Roberts et al., 2017; Oliveira et al., 2019). Depending on how the cross-validation is performed, the obtained performance will be indicative of one of these two dimensions. As it is shown in Sect. 3.2.3, the ANNs and especially the linear modelling approaches had more success in predicting SMB values in space than in time. This is mostly due to the fact that the glacier-wide SMB signal has a greater variability and nonlinearities in time than in space, with climate being the main driver of the annual fluctuations in SMB, whereas geography, and in particular the local topography, modulates the signal between glaciers ~~(Huss (2012); Rabatel et al. (2016); Vincent et al. (2017))~~(Huss, 2012; Rabatel et al., 2016; Vincent et al., 2017). Consequently, linear models find it easier to make predictions on a given period of time for other glaciers elsewhere in space, than for time periods outside the training. Nonetheless, the deep learning SMB models were capable of equally capturing the complex nonlinear patterns in both the spatial and temporal dimensions.

In order to cope with the specific challenges related to each type of cross-validation, there are several ~~parameters~~ hyperparameters that can be modified to adapt the ANN's behaviour. ~~The ANN~~Due to the long list of hyperparameters intervening in an ANN, it is not advisable to select them using brute force with a grid search or cross-validation. Instead, initial tests are performed in a subset of random folds to narrow down the range of best performing values, before moving to the full final cross-validations for the final hyperparameter selection. Moreover, the ANN architecture plays an important role: the number of neurons as well as the number of hidden layers will determine the ANN's complexity and its capabilities to capture hidden patterns in the data. But the larger the architecture, the higher are the chances to overfit the data. This undesired effect can be counterbalanced using regularization. The amount of regularization (dropout and Gaussian noise in our case, see Sect. 2.2.4) used in the training of the ANN necessarily introduces some trade-offs. The greater the dropout, the more we will constrain the learning of the ANN so the higher the generalization will be, until a certain point, where relevant information will start to be lost and performance will drop. On the other hand, the learning rate to compute the stochastic gradient descent, which tries to minimize the loss function, also plays an important role: smaller learning rates generally result in a slower convergence towards the absolute minima, thus producing models with better generalization. By balancing all these different effects, one can achieve the accuracy versus generalization ratio that best suits a certain dataset and model in terms of performance. Nonetheless, one key aspect in machine learning models is data: expanding the training dataset in the future will allow to increase the complexity of the model and its performance. Consequently, machine learning models see their performance improved as time goes by, with new data becoming available for training.

Although the features used as input for the model are classical descriptors of the topographical and meteorological conditions of the glaciers, it is worth mentioning that applying the model in different areas or with different data sources would likely require a re-training of the model due to possible biases: different regions on the globe may have other descriptors of importance but also different measuring techniques will likely have different biases.

4.3 Perspectives on future applications of deep learning in glaciology

The currently used meteorological variables in the deep ANN of ALPGM's SMB component are based on the classic degree-day approach, which relies only on temperature and precipitation. However, the model could be trained with variables involved in more complex models, such as SEB-type models, for which the longwave and shortwave radiation, as well as the turbulent fluxes and albedo intervene. The current model framework allows flexibility in the choice and number of input variables that can reflect different degrees of complexity for the resolved processes. Despite the fact that it has been shown that for glaciers in the European Alps there is almost no added value in transitioning from a simple degree-day to a SEB model for annual glacier-wide SMB simulations ~~(*e.g.*, Réveillet et al. (2017)),~~ (e.g., Réveillet et al., 2017), it could be an interesting way to expand the training dataset for glaciers in tropical and subtropical regions, where shortwave radiation plays a much more important role ~~(Benn and Evans (2014)). Maussion et al. (2015)~~ (Benn and Evans, 2014). Maussion et al. (2015) followed a similar approach with linear machine learning in order to calibrate a regression-based downscaling model that linked local SEB/SMB fluxes to atmospheric reanalysis variables. For ALPGM, the ~~ANN was trained with~~ SMB machine learning models were trained using glacier-wide SMB data, ~~but the~~ due to the high availability of glacier-wide SMB data in the French Alps (Rabatel et al., 2016). Nevertheless, the same approach could be used for point SMB data from field observations.

In this work, we also evaluated the resilience of the deep learning approach: since many ~~glaciarized~~ glacierized regions in the world do not have the same amount of data used in this study, we trained an ANN only with monthly average temperature and snowfall, without any topographical predictors, to see until which point the algorithm is capable of learning from minimal data. The results were quite ~~astonishing~~interesting, with a coefficient of determination of 0.68 (against 0.76 from the full model) and a RMSE of 0.59 (against 0.51 from the full model). These results indicate that meteorological data is the primary source of information~~regarding the annual~~, determining the interannual high frequency variability of the glacier-wide SMB ~~modelling, and~~ signal. On the other hand, the “bonus” of topographical data helps to modulate the high frequency climate signal, by adding a low frequency component to better differentiate glaciers ~~whose specificities do not stand out by just using climate data~~ and the topographical characteristics included in the glacier-wide SMB data (Huss et al., 2012). Consequently, the potential of deep learning in glacier modelling should not be ignored. A nonlinear deep learning SMB component like the one used for ALPGM could provide an interesting alternative to classical SMB models used for regional modelling. The comparison with other SMB models is beyond the scope of this study, but it would be worth investigating to quantify the specific gains that could be achieved by switching to a deep learning modelling approach. Nonetheless, the linear machine learning models trained with the CPDD and cumulative snowfall used in this study behave in a similar way to a calibrated temperature-index model.

Deep learning can be of special interest once applied in the reconstruction of SMB time series. More and more SMB data is becoming available thanks to the advances in remote sensing ~~(e.g., Brun et al. (2017); Rabatel et al. (2017); Zemp et al. (2019),~~ (e.g., Brun et al., 2017; Zemp et al., 2019; Dussaillant et al., 2019), but these datasets often cover limited areas and the most recent time period in the studied regions. An interesting way of expanding a dataset would be to use a deep learning approach to fill the data gaps~~. SMB past~~, based on the relationships found in a subset of glaciers as in the case study presented here. Past SMB time series of vast glaciarized regions could thereby be ~~obtained with unprecedented efficiency. Such an approach would be an excellent way of obtaining more SMB data~~ reconstructed, with potential applications in remote glaciarized regions such as the Andes or the Himalayas. ~~It could also be applied in data-rich regions benefiting from regionalized climate reanalyses (*e.g.*, Caillouet et al. (2016), covering the 1871-present period for France). Another possibility would be to completely bypass both the SMB and glacier dynamics of a classic glacier evolution model by training a deep ANN which would directly simulate changes in glacier thicknesses. If the ANN is trained with enough glacier thickness changes, climatic and topographical data, it could be able to simulate the 3D evolution of the glacier straight from the raw data. It might still be too soon for such models to be implemented, but once enough data will be available in the future, this could be a promising new way of tackling glacier evolution modelling.~~

~~Finally, yet another research perspective for deep learning and ANNs in glaciology is the estimation of glacier ice thickness. Clarke et al. (2009) already took the first steps towards this direction, but as in Steiner et al. (2005), they used a rather simple ANN, with sigmoid and hyperbolic tangent activation functions and just two hidden layers. Glacier ice thickness estimation is probably a much more complex problem than SMB simulation, so a more complex deep ANN fed by the Glacier Thickness Database (Consortium (2019)) could provide an interesting alternative to physical approaches.~~

5 Conclusions

~~We presented ALPGM (Bolibar (2019))~~We presented a novel approach to simulate and reconstruct glacier-wide SMB series using deep learning for individual glaciers at regional scale. This method has been included as a SMB component in ALPGM (Bolibar, 2019), a parameterized regional glacier evolution model, following an alternative approach to most physical and process-based glacier models. ~~Annual~~ The data-driven glacier-wide ~~SMBs are simulated using nonlinear deep learning, and the glacier geometry is updated annually using a~~ SMB modelling component is coupled with a glacier geometry update component, based on glacier-specific ~~parameterization~~parameterized functions. Deep learning is shown to outperform linear methods for the simulation of glacier-wide SMB with a case study of French alpine glaciers. By means of cross-validation, we demonstrated how important nonlinear structures (up to 35%) coming from the glacier and climate systems in both the spatial and temporal dimensions are captured by the deep ANN. Taking into account this nonlinearity substantially improved the explained variance and accuracy compared to linear statistical models, especially in the more complex temporal dimension. As we have shown in our case study, deep ANNs are capable of dealing with relatively small datasets, and they present a wide range of configurations to generalize and prevent overfitting. Machine learning models benefit from the increasing number of available data, which makes their performance constantly improve as time goes by.

Deep learning should be seen as an opportunity by the glaciology community. Its good performance for SMB modelling in both the spatial and temporal dimensions shows how relevant it can be for a broad range of applications. Combined with in situ or remote sensing SMB estimations, it can serve to ~~extend~~ reconstruct SMB time series for regions or glaciers with already available data for past and future periods~~; or it could help extend SMB datasets to other unmeasured glaciers for remote regions with only few monitored glaciers ,~~, with potential applications in remote regions such as the Andes or the high mountains of Asia. Moreover, deep learning ~~could~~ can be used as an alternative to classical SMB models ~~;~~as it is done in ALPGM: important nonlinearities from the glacier and climate systems are potentially ignored by these mostly linear models, which could give an advantage to deep learning models in regional studies. It might still be too early for the development of such models in certain regions ~~, but~~ which lack consistent datasets with a good spatial and temporal coverage. Nevertheless, as new data becomes available the gap is slowly being closed towards real big data approaches in glaciology.

Code and data availability. The source code of ALPGM is available at https://github.com/JordiBolibar/ALPGM with its DOI (10.5281/zenodo.3269678) for the v1.0 release. All scripts used to generate plots and results are included in the repository. The detailed information regarding glaciers used in the case study and the SMB model performance per glacier are included in the Supplementary material.

Author contributions. J. Bolibar developed ALPGM, analysed the results and wrote the paper. A. Rabatel did the glaciological analysis, contributed with the remote sensing SMB data, and conceived the study together with I. Gouttevin, T. Condom and J. Bolibar. C. Galiez and J. Bolibar developed the deep and machine learning modelling approach, and E. Sauquet contributed in the statistical analysis. All authors discussed the results and helped to develop the paper.

Competing interests. The authors declare that they do not have any competing interests.

Acknowledgements. This study has been possible through funding from the ANR VIP_Mont-Blanc (ANR-14 CE03-0006-03), the BERGER project (Pack Ambition Recherche funded by the AuRA region), the LabEx OSUG@2020 (Investissements d'avenir - ANR10 LABX56) and the CNES, via the KALEIDOS-Alpes and ISIS projects. Data has been provided by Météo-France (SAFRAN ~~and ADAMONT datasets~~dataset) and the GLACIOCLIM National Observation Service (*in situ* glaciological data). The authors would like to thank ~~Benjamin Renard (Irstea Lyon) for his insightful comments on statistics and~~ Eduardo Pérez-Pellitero (Max Planck Institute for Intelligent Systems) for the interesting discussions regarding deep learning, Benjamin Renard (Irstea Lyon) for his insightful comments on statistics, as well as Fabien Maussion and an anonymous reviewer for their constructive review comments which have helped to improve the overall quality and clarity of the manuscript.

References

Beniston, M., Farinotti, D., Stoffel, M., Andreassen, L. M., Coppola, E., Eckert, N., Fantini, A., Giacona, F., Hauck, C., Huss, M., Huwald, H., Lehning, M., López-Moreno, J.-I., Magnusson, J., Marty, C., Morán-Tejéda, E., Morin, S., Naaim, M., Provenzale, A., Rabatel, A., Six, D., Stötter, J., Strasser, U., Terzago, S., and Vincent, C.: The European mountain cryosphere: a review of its current state, trends, and future challenges, The Cryosphere, 12, 759–794, https://doi.org/10.5194/tc-12-759-2018, https://www.the-cryosphere.net/12/759/2018/, 2018.

Benn, D. I. and Evans, D. J. A.: Glaciers & glaciation, Routledge, New York, NY, USA, 2nd edn., http://www.imperial.eblib.com/EBLWeb/patron/?target=patron&extendedid=P_615876_0, oCLC: 878863282, 2014.

Bolibar, J.: JordiBolibar/ALPGM: ALPGM v1.0, https://doi.org/10.5281/zenodo.3269678, https://zenodo.org/record/3269678, 2019.

Brun, F., Berthier, E., Wagnon, P., Kääb, A., and Treichler, D.: A spatially resolved estimate of High Mountain Asia glacier mass balances from 2000 to 2016, Nature Geoscience, 10, 668–673, https://doi.org/10.1038/ngeo2999, http://www.nature.com/doifinder/10.1038/ngeo2999, 2017.

Caillouet, L., Vidal, J.-P., Sauquet, E., and Graff, B.: Probabilistic precipitation and temperature downscaling of the Twentieth Century Reanalysis over France, Climate of the Past, 12, 635–662, https://doi.org/10.5194/cp-12-635-2016, https://www.clim-past.net/12/635/2016/, 2016.

Carlson, B. Z., Georges, D., Rabatel, A., Randin, C. F., Renaud, J., Delestrade, A., Zimmermann, N. E., Choler, P., and Thuiller, W.: Accounting for tree line shift, glacier retreat and primary succession in mountain plant distribution models, Diversity and Distributions, 20, 1379–1391, https://doi.org/10.1111/ddi.12238, http://doi.wiley.com/10.1111/ddi.12238, 2014.

Chollet, F.: Keras, https://keras.io, 2015.

Clarke, G. K. C., Berthier, E., Schoof, C. G., and Jarosch, A. H.: Neural Networks Applied to Estimating Subglacial Topography and Glacier Volume, Journal of Climate, 22, 2146–2160, https://doi.org/10.1175/2008JCLI2572.1, http://journals.ametsoc.org/doi/abs/10.1175/2008JCLI2572.1, 2009.

Consortium, G.: Glacier Thickness Database 3.0.1, 2019.

Consortium, R. G. I.: Randolph Glacier Inventory 6.0, https://doi.org/10.7265/N5-RGI-60, http://www.glims.org/RGI/randolph60.html, type: dataset, 2017.

Ducournau, A. and Fablet, R.: Deep learning for ocean remote sensing: an application of convolutional neural networks for super-resolution on satellite-derived SST data, in: 2016 9th IAPR Workshop on Pattern Recogniton in Remote Sensing (PRRS), pp. 1–6, IEEE, Cancun, Mexico, https://doi.org/10.1109/PRRS.2016.7867019, http://ieeexplore.ieee.org/document/7867019/, 2016.

Durand, Y., Laternser, M., Giraud, G., Etchevers, P., Lesaffre, B., and Mérindol, L.: Reanalysis of 44 Yr of Climate in the French Alps (1958–2002): Methodology, Model Validation, Climatology, and Trends for Air Temperature and Precipitation, Journal of Applied Meteorology and Climatology, 48, 429–449, https://doi.org/10.1175/2008JAMC1808.1, http://journals.ametsoc.org/doi/abs/10.1175/2008JAMC1808.1, 2009.

Dussaillant, I., Berthier, E., Brun, F., Masiokas, M., Hugonnet, R., Favier, V., Rabatel, A., Pitte, P., and Ruiz, L.: Two decades of glacier mass loss along the Andes, Nature Geoscience, https://doi.org/10.1038/s41561-019-0432-5, http://www.nature.com/articles/s41561-019-0432-5, 2019.

Farinotti, D., Huss, M., Fürst, J. J., Landmann, J., Machguth, H., Maussion, F., and Pandit, A.: A consensus estimate for the ice thickness distribution of all glaciers on Earth, Nature Geoscience, 12, 168–173, https://doi.org/10.1038/s41561-019-0300-3, http://www.nature.com/articles/s41561-019-0300-3, 2019.

Fausett, L. V.: Fundamentals of neural networks: architectures, algorithms, and applications, Prentice Hall, Englewood Cliffs, N.J., oCLC: 28215780, 1994.

Gagliardini, O., Zwinger, T., Gillet-Chaulet, F., Durand, G., Favier, L., de Fleurian, B., Greve, R., Malinen, M., Martín, C., Råback, P., Ruokolainen, J., Sacchettini, M., Schäfer, M., Seddik, H., and Thies, J.: Capabilities and performance of Elmer/Ice, a new-generation ice sheet model, Geoscientific Model Development, 6, 1299–1318, https://doi.org/10.5194/gmd-6-1299-2013, https://www.geosci-model-dev.net/6/1299/2013/, 2013.

Gardent, M., Rabatel, A., Dedieu, J.-P., and Deline, P.: Multitemporal glacier inventory of the French Alps from the late 1960s to the late 2000s, Global and Planetary Change, 120, 24–37, https://doi.org/10.1016/j.gloplacha.2014.05.004, https://linkinghub.elsevier.com/retrieve/pii/S092181811400099X, 2014.

Gerbaux, M., Genthon, C., Etchevers, P., Vincent, C., and Dedieu, J.: Surface mass balance of glaciers in the French Alps: distributed modeling and sensitivity to climate change, Journal of Glaciology, 51, 561–572, https://doi.org/10.3189/172756505781829133, https://www.cambridge.org/core/product/identifier/S0022143000210769/type/journal_article, 2005.

Hanzer, F., Förster, K., Nemec, J., and Strasser, U.: Projected cryospheric and hydrological impacts of 21st century climate change in the Ötztal Alps (Austria) simulated using a physically based approach, Hydrology and Earth System Sciences, 22, 1593–1614, https://doi.org/10.5194/hess-22-1593-2018, https://www.hydrol-earth-syst-sci.net/22/1593/2018/, 2018.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer Series in Statistics, Springer New York, New York, NY, https://doi.org/10.1007/978-0-387-84858-7, http://link.springer.com/10.1007/978-0-387-84858-7, 2009.

Hawkins, D. M.: The Problem of Overfitting, Journal of Chemical Information and Computer Sciences, 44, 1–12, https://doi.org/10.1021/ci0342472, http://pubs.acs.org/doi/abs/10.1021/ci0342472, 2004.

He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, 2015 IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/iccv.2015.123, http://dx.doi.org/10.1109/ICCV.2015.123, 2015.

Hock, R.: Temperature index melt modelling in mountain areas, Journal of Hydrology, 282, 104–115, https://doi.org/10.1016/S0022-1694(03)00257-9, http://linkinghub.elsevier.com/retrieve/pii/S0022169403002579, 2003.

Hock, R., Bliss, A., Marzeion, B., Giesen, R. H., Hirabayashi, Y., Huss, M., Radić, V., and Slangen, A. B. A.: GlacierMIP – A model intercomparison of global-scale glacier mass-balance models and projections, Journal of Glaciology, 65, 453–467, https://doi.org/10.1017/jog.2019.22, https://www.cambridge.org/core/product/identifier/S0022143019000224/type/journal_article, 2019.

Hoerl, A. E. and Kennard, R. W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, 12, 55–67, https://doi.org/10.1080/00401706.1970.10488634, http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634, 1970.

Hoinkes, H. C.: Glacier Variation and Weather, Journal of Glaciology, 7, 3–18, https://doi.org/10.3189/S0022143000020384, https://www.cambridge.org/core/product/identifier/S0022143000020384/type/journal_article, 1968.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Dai, X., Maskell, K., and Johnson, C.: Climate change 2001: the scientific basis, The Press Syndicate of the University of Cambridge, 2001.

Huss, M.: Extrapolating glacier mass balance to the mountain-range scale: the European Alps 1900–2100, The Cryosphere, 6, 713–727, https://doi.org/10.5194/tc-6-713-2012, https://www.the-cryosphere.net/6/713/2012/, 2012.

Huss, M. and Hock, R.: A new model for global glacier change and sea-level rise, Frontiers in Earth Science, 3, https://doi.org/10.3389/feart.2015.00054, http://journal.frontiersin.org/Article/10.3389/feart.2015.00054/abstract, 2015.

Huss, M. and Hock, R.: Global-scale hydrological response to future glacier mass loss, Nature Climate Change, 8, 135–140, https://doi.org/10.1038/s41558-017-0049-x, http://www.nature.com/articles/s41558-017-0049-x, 2018.

Huss, M., Farinotti, D., Bauder, A., and Funk, M.: Modelling runoff from highly glacierized alpine drainage basins in a changing climate, Hydrological Processes, 22, 3888–3902, https://doi.org/10.1002/hyp.7055, http://doi.wiley.com/10.1002/hyp.7055, 2008.

Huss, M., Jouvet, G., Farinotti, D., and Bauder, A.: Future high-mountain hydrology: a new parameterization of glacier retreat, Hydrology and Earth System Sciences, 14, 815–829, https://doi.org/10.5194/hess-14-815-2010, http://www.hydrol-earth-syst-sci.net/14/815/2010/, 2010.

Huss, M., Hock, R., Bauder, A., and Funk, M.: Conventional versus reference-surface mass balance, Journal of Glaciology, 58, 278–286, https://doi.org/10.3189/2012JoG11J216, https://www.cambridge.org/core/product/identifier/S0022143000212021/type/journal_article, 2012.

Ingrassia, S. and Morlini, I.: Neural Network Modeling for Small Datasets, Technometrics, 47, 297–311, https://doi.org/10.1198/004017005000000058, http://www.tandfonline.com/doi/abs/10.1198/004017005000000058, 2005.

Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, 2015.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 2018.

Jiang, G.-Q., Xu, J., and Wei, J.: A Deep Learning Algorithm of Neural Network for the Parameterization of Typhoon-Ocean Feedback in Typhoon Forecast Models, Geophysical Research Letters, 45, 3706–3716, https://doi.org/10.1002/2018GL077004, http://doi.wiley.com/10.1002/2018GL077004, 2018.

Jouvet, G., Huss, M., Blatter, H., Picasso, M., and Rappaz, J.: Numerical simulation of Rhonegletscher from 1874 to 2100, Journal of Computational Physics, 228, 6426–6439, https://doi.org/10.1016/j.jcp.2009.05.033, https://linkinghub.elsevier.com/retrieve/pii/S002199910900285X, 2009.

Jóhannesson, T., Raymond, C., and Waddington, E.: Time–Scale for Adjustment of Glaciers to Changes in Mass Balance, Journal of Glaciology, 35, 355–369, https://doi.org/10.3189/S002214300000928X, https://www.cambridge.org/core/product/identifier/S002214300000928X/type/journal_article, 1989.

Lguensat, R., Sun, M., Fablet, R., Tandeo, P., Mason, E., and Chen, G.: EddyNet: A Deep Neural Network For Pixel-Wise Classification of Oceanic Eddies, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 1764–1767, IEEE, Valencia, https://doi.org/10.1109/IGARSS.2018.8518411, https://ieeexplore.ieee.org/document/8518411/, 2018.

Martin, S.: Correlation bilans de masse annuels-facteurs météorologiques dans les Grandes Rousses, Zeitschrift für Gletscherkunde und Glazialgeologie, 1974.

Marzeion, B., Jarosch, A. H., and Hofer, M.: Past and future sea-level change from the surface mass balance of glaciers, The Cryosphere, 6, 1295–1322, https://doi.org/10.5194/tc-6-1295-2012, https://www.the-cryosphere.net/6/1295/2012/, 2012.

Marçais, J. and de Dreuzy, J.-R.: Prospective Interest of Deep Learning for Hydrological Inference: J. Marçais and J.-R. de Dreuzy Groundwater xx, no. x: xx-xx, Groundwater, 55, 688–692, https://doi.org/10.1111/gwat.12557, http://doi.wiley.com/10.1111/gwat.12557, 2017.

Maussion, F., Gurgiser, W., Großhauser, M., Kaser, G., and Marzeion, B.: ENSO influence on surface energy and mass balance at Shallap Glacier, Cordillera Blanca, Peru, The Cryosphere, 9, 1663–1683, https://doi.org/10.5194/tc-9-1663-2015, https://www.the-cryosphere.net/9/1663/2015/, 2015.

Maussion, F., Butenko, A., Champollion, N., Dusch, M., Eis, J., Fourteau, K., Gregor, P., Jarosch, A. H., Landmann, J., Oesterle, F., Recinos, B., Rothenpieler, T., Vlug, A., Wild, C. T., and Marzeion, B.: The Open Global Glacier Model (OGGM) v1.1, Geoscientific Model Development, 12, 909–931, https://doi.org/10.5194/gmd-12-909-2019, https://www.geosci-model-dev.net/12/909/2019/, 2019.

NSIDC, G. a.: Global Land Ice Measurements from Space glacier database. Compiled and made available by the international GLIMS community and the National Snow and Ice Data Center, 2005.

Nussbaumer, S., Steiner, D., and Zumbühl, H.: Réseau neuronal et fluctuations des glaciers dans les Alpes occidentales, 2012.

Oliveira, M., Torgo, L., and Santos Costa, V.: Evaluation Procedures for Forecasting with Spatio-Temporal Data, in: Machine Learning and Knowledge Discovery in Databases, edited by Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., and Ifrim, G., vol. 11051, pp. 703–718, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-10925-7_43, http://link.springer.com/10.1007/978-3-030-10925-7_43, 2019.

Olson, M., Wyner, A. J., and Berk, R.: Modern Neural Networks Generalize on Small Data Sets, in: NeurIPS, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., and Louppe, G.: Scikit-learn: Machine Learning in Python, vol. 12, 2012.

Rabatel, A., Letréguilly, A., Dedieu, J.-P., and Eckert, N.: Changes in glacier equilibrium-line altitude in the western Alps from 1984 to 2010: evaluation by remote sensing and modeling of the morpho-topographic and climate controls, The Cryosphere, 7, 1455–1471, https://doi.org/10.5194/tc-7-1455-2013, https://www.the-cryosphere.net/7/1455/2013/, 2013.

Rabatel, A., Dedieu, J. P., and Vincent, C.: Spatio-temporal changes in glacier-wide mass balance quantified by optical remote sensing on 30 glaciers in the French Alps for the period 1983–2014, Journal of Glaciology, 62, 1153–1166, https://doi.org/10.1017/jog.2016.113, https://www.cambridge.org/core/product/identifier/S0022143016001131/type/journal_article, 2016.

Rabatel, A., Sirguey, P., Drolon, V., Maisongrande, P., Arnaud, Y., Berthier, E., Davaze, L., Dedieu, J.-P., and Dumont, M.: Annual and Seasonal Glacier-Wide Surface Mass Balance Quantified from Changes in Glacier Surface State: A Review on Existing Methods Using Optical Satellite Imagery, Remote Sensing, 9, 507, https://doi.org/10.3390/rs9050507, http://www.mdpi.com/2072-4292/9/5/507, 2017.

Rabatel, A., Sanchez, O., Vincent, C., and Six, D.: Estimation of Glacier Thickness From Surface Mass Balance and Ice Flow Velocities: A Case Study on Argentière Glacier, France, Frontiers in Earth Science, 6, https://doi.org/10.3389/feart.2018.00112, https://www.frontiersin.org/article/10.3389/feart.2018.00112/full, 2018.

Radić, V., Bliss, A., Beedlow, A. C., Hock, R., Miles, E., and Cogley, J. G.: Regional and global projections of twenty-first century glacier mass changes in response to climate scenarios from global climate models, Climate Dynamics, 42, 37–58, https://doi.org/10.1007/s00382-013-1719-7, http://link.springer.com/10.1007/s00382-013-1719-7, 2014.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, http://www.pnas.org/lookup/doi/10.1073/pnas.1810286115, 2018.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, http://doi.wiley.com/10.1111/ecog.02881, 2017.

Réveillet, M., Rabatel, A., Gillet-Chaulet, F., and Soruco, A.: Simulations of changes to Glaciar Zongo, Bolivia (16° S), over the 21st century using a 3-D full-Stokes model and CMIP5 climate projections, Annals of Glaciology, 56, 89–97, https://doi.org/10.3189/2015AoG70A113, https://www.cambridge.org/core/product/identifier/S0260305500250362/type/journal_article, 2015.

Réveillet, M., Vincent, C., Six, D., and Rabatel, A.: Which empirical model is best suited to simulate glacier mass balances?, Journal of Glaciology, 63, 39–54, https://doi.org/10.1017/jog.2016.110, https://www.cambridge.org/core/product/identifier/S0022143016001106/type/journal_article, 2017.

Réveillet, M., Six, D., Vincent, C., Rabatel, A., Dumont, M., Lafaysse, M., Morin, S., Vionnet, V., and Litt, M.: Relative performance of empirical and physical models in assessing the seasonal and annual glacier surface mass balance of Saint-Sorlin Glacier (French Alps), The Cryosphere, 12, 1367–1386, https://doi.org/10.5194/tc-12-1367-2018, https://www.the-cryosphere.net/12/1367/2018/, 2018.

Seabold, S. and Perktold, J.: Statsmodels: Econometric and Statistical Modelingwith Python, PROC. OF THE 9th PYTHON IN SCIENCE CONF., 2010.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resources Research, 54, 8558–8593, https://doi.org/10.1029/2018WR022643, https://onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022643, 2018.

Six, D. and Vincent, C.: Sensitivity of mass balance and equilibrium-line altitude to climate change in the French Alps, Journal of Glaciology, 60, 867–878, https://doi.org/10.3189/2014JoG14J014, https://www.cambridge.org/core/product/identifier/S0022143000202128/type/journal_article, 2014.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1929–1958, 2014.

Steiner, D., Walter, A., and Zumbühl, H.: The application of a non-linear back-propagation neural network to study the mass balance of Grosse Aletschgletscher, Switzerland, Journal of Glaciology, 51, 313–323, https://doi.org/10.3189/172756505781829421, https://www.cambridge.org/core/product/identifier/S0022143000214949/type/journal_article, 2005.

Steiner, D., Pauling, A., Nussbaumer, S. U., Nesje, A., Luterbacher, J., Wanner, H., and Zumbühl, H. J.: Sensitivity of European glaciers to precipitation and temperature – two case studies, Climatic Change, 90, 413–441, https://doi.org/10.1007/s10584-008-9393-1, http://link.springer.com/10.1007/s10584-008-9393-1, 2008.

Thibert, E., Dkengne Sielenou, P., Vionnet, V., Eckert, N., and Vincent, C.: Causes of Glacier Melt Extremes in the Alps Since 1949, Geophysical Research Letters, 45, 817–825, https://doi.org/10.1002/2017GL076333, http://doi.wiley.com/10.1002/2017GL076333, 2018.

Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological), 58, 267–288, http://www.jstor.org/stable/2346178, 1996.

Tibshirani, R., Johnstone, I., Hastie, T., and Efron, B.: Least angle regression, The Annals of Statistics, 32, 407–499, https://doi.org/10.1214/009053604000000067, http://projecteuclid.org/euclid.aos/1083178935, 2004.

Vincent, C., Harter, M., Gilbert, A., Berthier, E., and Six, D.: Future fluctuations of Mer de Glace, French Alps, assessed using a parameterized model calibrated with past thickness changes, Annals of Glaciology, 55, 15–24, https://doi.org/10.3189/2014AoG66A050, https://www.cambridge.org/core/product/identifier/S0260305500258096/type/journal_article, 2014.

Vincent, C., Fischer, A., Mayer, C., Bauder, A., Galos, S. P., Funk, M., Thibert, E., Six, D., Braun, L., and Huss, M.: Common climatic signal from glaciers in the European Alps over the last 50 years: Common Climatic Signal in the Alps, Geophysical Research Letters, 44, 1376–1383, https://doi.org/10.1002/2016GL072094, http://doi.wiley.com/10.1002/2016GL072094, 2017.

Vincent, C., Peyaud, V., Laarman, O., Six, D., Gilbert, A., Gillet-Chaulet, F., Berthier, , Morin, S., Verfaillie, D., Rabatel, A., Jourdain, B., and Bolibar, J.: Déclin des deux plus grands glaciers des Alpes françaises au cours du XXIe siècle : Argentière et Mer de Glace, La Météorologie, p. 49, https://doi.org/10.4267/2042/70369, http://hdl.handle.net/2042/70369, 2019.

Vionnet, V., Dombrowski-Etchevers, I., Lafaysse, M., Quéno, L., Seity, Y., and Bazile, E.: Numerical Weather Forecasts at Kilometer Scale in the French Alps: Evaluation and Application for Snowpack Modeling, Journal of Hydrometeorology, 17, 2591–2614, https://doi.org/10.1175/JHM-D-15-0241.1, http://journals.ametsoc.org/doi/10.1175/JHM-D-15-0241.1, 2016.

Vuille, M., Carey, M., Huggel, C., Buytaert, W., Rabatel, A., Jacobsen, D., Soruco, A., Villacis, M., Yarleque, C., Elison Timm, O., Condom, T., Salzmann, N., and Sicart, J.-E.: Rapid decline of snow and ice in the tropical Andes – Impacts, uncertainties and challenges ahead, Earth-Science Reviews, 176, 195–213, https://doi.org/10.1016/j.earscirev.2017.09.019, https://linkinghub.elsevier.com/retrieve/pii/S0012825216304512, 2018.

Weisberg, S.: Applied linear regression, Wiley series in probability and statistics, Wiley, Hoboken, NJ, fourth edition edn., 2014.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., and Freckleton, R. P.: Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour, Journal of Animal Ecology, 75, 1182–1189, https://doi.org/10.1111/j.1365-2656.2006.01141.x, http://doi.wiley.com/10.1111/j.1365-2656.2006.01141.x, 2006.

Xu, B., Wang, N., Chen, T., and Li, M.: Empirical Evaluation of Rectified Activations in Convolutional Network, CoRR, abs/1505.00853, http://arxiv.org/abs/1505.00853, 2015.

Zekollari, H. and Huybrechts, P.: Statistical modelling of the surface mass-balance variability of the Morteratsch glacier, Switzerland: strong control of early melting season meteorological conditions, Journal of Glaciology, 64, 275–288, https://doi.org/10.1017/jog.2018.18, https://www.cambridge.org/core/product/identifier/S0022143018000187/type/journal_article, 2018.

Zekollari, H., Huss, M., and Farinotti, D.: Modelling the future evolution of glaciers in the European Alps under the EURO-CORDEX RCM ensemble, The Cryosphere, 13, 1125–1146, https://doi.org/10.5194/tc-13-1125-2019, https://www.the-cryosphere.net/13/1125/2019/, 2019.

Zemp, M., Haeberli, W., Hoelzle, M., and Paul, F.: Alpine glaciers to disappear within decades?, Geophysical Research Letters, 33, https://doi.org/10.1029/2006GL026319, http://doi.wiley.com/10.1029/2006GL026319, 2006.

Zemp, M., Huss, M., Thibert, E., Eckert, N., McNabb, R., Huber, J., Barandun, M., Machguth, H., Nussbaumer, S. U., Gärtner-Roer, I., Thomson, L., Paul, F., Maussion, F., Kutuzov, S., and Cogley, J. G.: Global glacier mass changes and their contributions to sea-level rise from 1961 to 2016, Nature, 568, 382–386, https://doi.org/10.1038/s41586-019-1071-0, http://www.nature.com/articles/s41586-019-1071-0, 2019.