Authors reply to Anonymous Referee #1’s review on “Deep learning applied to glacier evolution modelling”

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.

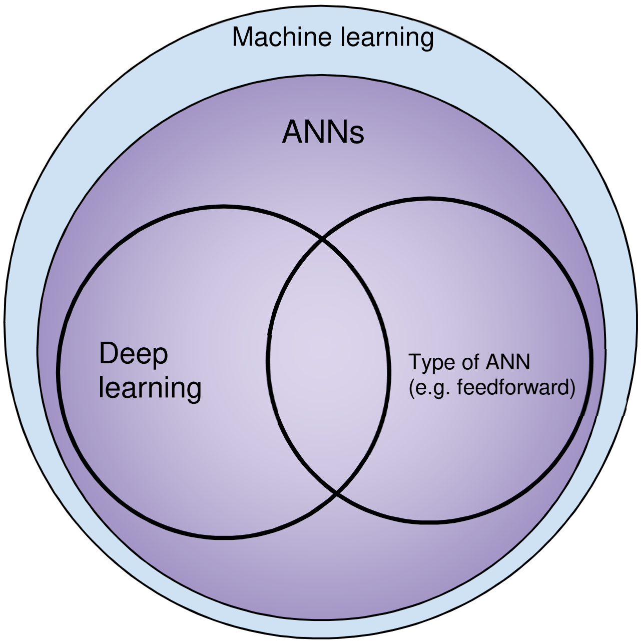
We would like to thank the reviewer for the time dedicated to read the manuscript and for the overall positive feedback. All the detailed comments will hopefully improve the overall clarity and fluidity of the manuscript. All points raised during the review have been addressed and answered, in the following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

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.**"

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:

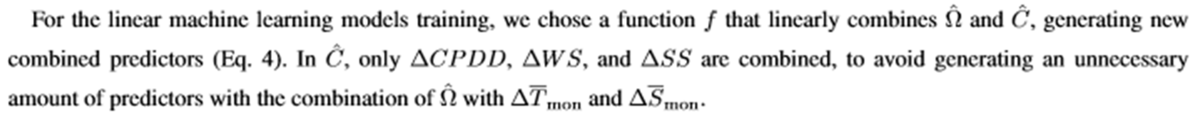
“For the linear machine learning model**s** training, we chose a function f that …”

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.

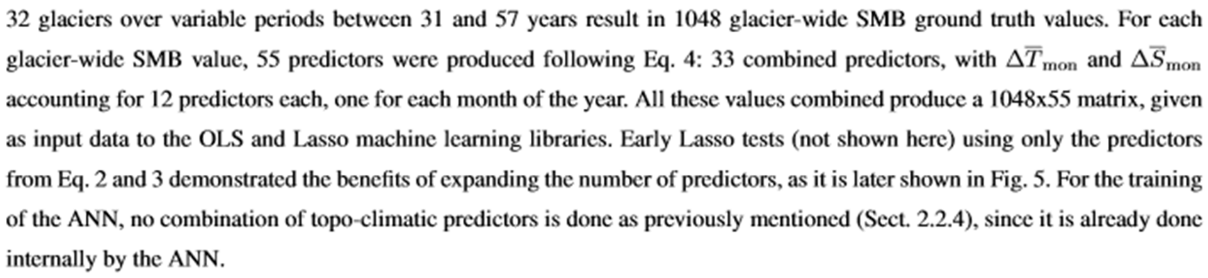
Page 16, Line 1: It is not clear to me why there are 50 predictors, when there are 33 coefficients in Equation 4.

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.

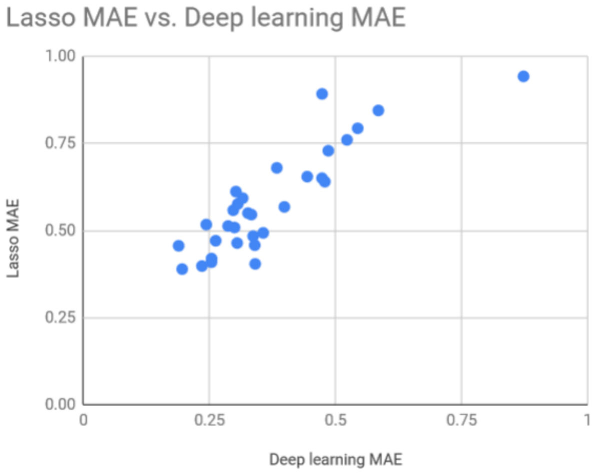
Page 16, Line 8: "...the annual CPDD as well as the winter and summer snowfall appear as significant predictors as well as several monthly mean temperatures and snowfall values..."; Change to "...the annual CPDD, winter and summer snowfall, and several monthly mean temperature and snowfall were found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

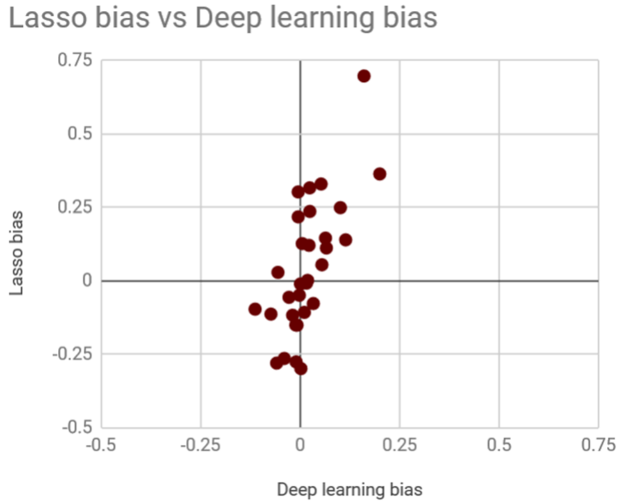
In this sentence we did not imply to use the word "significant" in statistical terms. A more accurate word choice, in the context of the causal analysis which determines the importance (%) of each predictor would be:

"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.