Authors reply to Ben Marzeion on "A deep learning reconstruction of mass balance series for all glaciers in the French Alps: 1967–2015"

Ben Marzeion

1 General comments

Bolibar et al. present the results of a new approach to reconstruct glacier mass balances at times and/or locations where meteorological conditions (and some topographical information) are known, but no observations of glacier mass balance exist. Their approach, based on a neural network algorithm, adds considerable diversity to the existing group of reconstruction methods. The thorough validation of the results leads to great confidence in the robustness of the method. Except for some minor issue listed below, the manuscript is very clear and easy to follow. The data set produced and presented here will be of great use for the community. I particularly appreciate the great care that has been taken in documenting the test for overfitting in the supplementary material. I recommend publication once the authors have gone through the list of questions/suggestions below.

We are grateful for the positive and encouraging comments. These comments will help improve the manuscript's quality and clarity. Most figures in the paper have been re-processed taking into account the feedback, hopefully leading to better visualization and presentation. Every comment/suggestion has been addressed individually in the following section.

As explained in one of the comments regarding the validation approach based on cross-validation, we have trained a new cross-validation ensemble of 60 members and updated the dataset results. This new ensemble is based on weighted bagging (Hastie et al., 2009) of Leave-Some-Years-and-Glaciers-Out cross-validation (Bolibar et al., 2020), which balances the training data in the model in order to better take into account the lack of data between 1967-1983. The main results and conclusions have not changed, only leading to a slightly less negative average mass balance (from -0.72 to -0.71 m w.e. a^{-1}), and slightly higher uncertainties due to the increased presence of underrepresented values of the 1967-1983 period (RMSE: from 0.49 to 0.55 m w.e. a^{-1} and r^2: from 0.79 to 0.75). We believe this even more rigorous cross-validation leads to more accurate results and uncertainty estimations.

2 Specific/minor comments

P1 L9: please specify "1\sigma" instead of "\sigma" for clarity.

The sentence in the abstract has been updated as suggested by the reviewer.

P1 L10: the "moderately" should only apply to the 1980s, I think

Indeed. The sentence has been updated:

“We estimate an average regional area-weighted glacier-wide SMB of -0.72±0.20 (1\sigma) m.w.e. a^{-1} for the 1967-2015 period, **with negative mass balances** in the 1970s (-0.52 m.w.e. a^{-1}), **moderately negative** in the 1980s (-0.12 m.w.e. a^{-1}), and an increasing negative trend from the 1990s onwards, up to -1.39 m.w.e. a^{-1} in the 2010s.”

P1 L10: avoid line break within negative number

This has been fixed with the rephrasing of some parts of the abstract.

P1 L12: unclear, what “this period” refers to

The sentence has been updated to clearly indicate the time period:

“Following a topographical and regional analysis, we estimate that the massifs with the highest mass losses **for the 1967-2015** period are the...”

abstract: why are no uncertainties given for the values of the different massifs? (also concerns the conclusions)

Because we have no way to dissociate the uncertainties for each massif from the overall uncertainties computed through cross-validation. Therefore, all massifs would display the same uncertainty, which is already given with the average performance of the method for this region (RMSE = 0.55 m w.e. a^{-1}). If the reviewer thinks it would still be better to give the uncertainty for each massif we can add it in the abstract and text.

P2 L8: “these points” refers to the points of MB measurements, but this reference is not very clear here; also, it’s not the points that show nonlinear variability, but the measurements at the points; suggest to rephrase

This sentence has been adapted to improve clarity as suggested:

“**These different point SMB measurements** can show a high nonlinear variability...”

P2 L23: there more four global parameters in the Marzeion et al. (2012) model, and I wouldn't necessarily say they were "optimized", because that "optimization" was very subjective...

The word "optimized" has been removed to erase these connotations from the sentence as suggested by the reviewer:

"They used a minimal model relying only on temperature and precipitation data, based on a temperature-index method, with two parameters to calibrate the temperature sensitivity and the precipitation lapse rate."

Fig. 1: the figure certainly works well for presentations etc., but I'm not sure it is necessary here, since the text describes very well what is done, and there is little to be gained from the figure.

Indeed, the main key aspects of the overall analysis are already given in the abstract and in the text. Nonetheless, we believe it is a complementary way to show the regional variability, as it shows in a single figure the spread of glacier behaviour and the common variability in a nice and easy way. This is our personal opinion, if the reviewer strongly suggests to remove it, we will move it to the supplementary material.

P3 L14-15: it would be great if you can add a sentence or two here, specifying how any difference in the altitude of the glaciers' centroids and the reanalysis grid points were treated (lapse rates or similar?)

The explanation on climate data and predictors has been updated with the following sentences in order to give some context on how the forcings are adjusted to each glacier centroid's altitude:

"(1) climate data from the SAFRAN meteorological reanalyses (Durand et al., 2009) with: cumulative positive degree days (CPDD), cumulative winter snowfall, cumulative summer snowfall, mean monthly temperature and mean monthly snowfall, all variables being quantified at the altitude of the glacier's centroid. **In order to capture the climate signal at each glacier's centroid, temperatures are taken from the nearest SAFRAN 300 m altitudinal band and adjusted with a 6ºC/km lapse rate. The updated temperature is then used to update the snowfall amount from the same 300 m altitudinal band.**"

P4 L22 or lower: It might be worth pointing out/discussing that the density of observations used in the LOGO cross validation is denser towards the end of the reconstruction interval, when presumably, also the quality of the meteorological data are higher, such that the uncertainty of the methods might be underestimated for the (roughly) first half of the period. I also wonder if/how this interferes with your assessment of the model's ability to reconstruct the more neutral MB values during 1967-1984?

That’s a very good point. That was one of our main concerns during the validation process, which we tried to address in two different ways.

First of all, we performed a separate cross-validation with only data from the 1967-1984 period, in order to specifically assess the performance during this period. This is explained in the newly created Sect. “2.3 Uncertainty assessment” (as suggested by Anonymous reviewer 1). This was already present in the version of the manuscript sent for review.

On the other hand, in order to improve our estimates and to better take into account this lack of homogeneity in the dataset, we have trained a new ensemble of models based on Leave-Some-Years-and-Glaciers-Out (LSYGO) cross-validation, as explained in Bolibar et al. (2020). We used an ensemble of 60 CV models using weighted bagging (Hastie et al., 2009) by giving +33% more weight to data between 1967-1984, in order to compensate for this lack of observations during this period, which covers a third of the 49-year period. This has not affected much the results, and the conclusions remain exactly the same, but it allows giving a more accurate and realistic assessment of the model’s performance, with a RMSE of 0.55 m w.e. a^{-1}, a coefficient of determination of 0.75 and an average bias of -0.019 m w.e. a^{-1}.

Fig. 2: since there are so many lines, it is somewhat hard to see the distribution. Particularly in the lower panel, a histogram for showing the distribution of the accumulated values (vertically, to the right of the panel) would be quite interesting. It would be possible to see, e.g., how/if the area weighted mean differs from the “ensemble” mean and/or median, if the distribution is (a)symmetric, etc. Just a suggestion to consider.

That is a good idea. A panel to the right of the cumulative plot has been added with a histogram, the PDF and the position of the area weighted mean SMB.

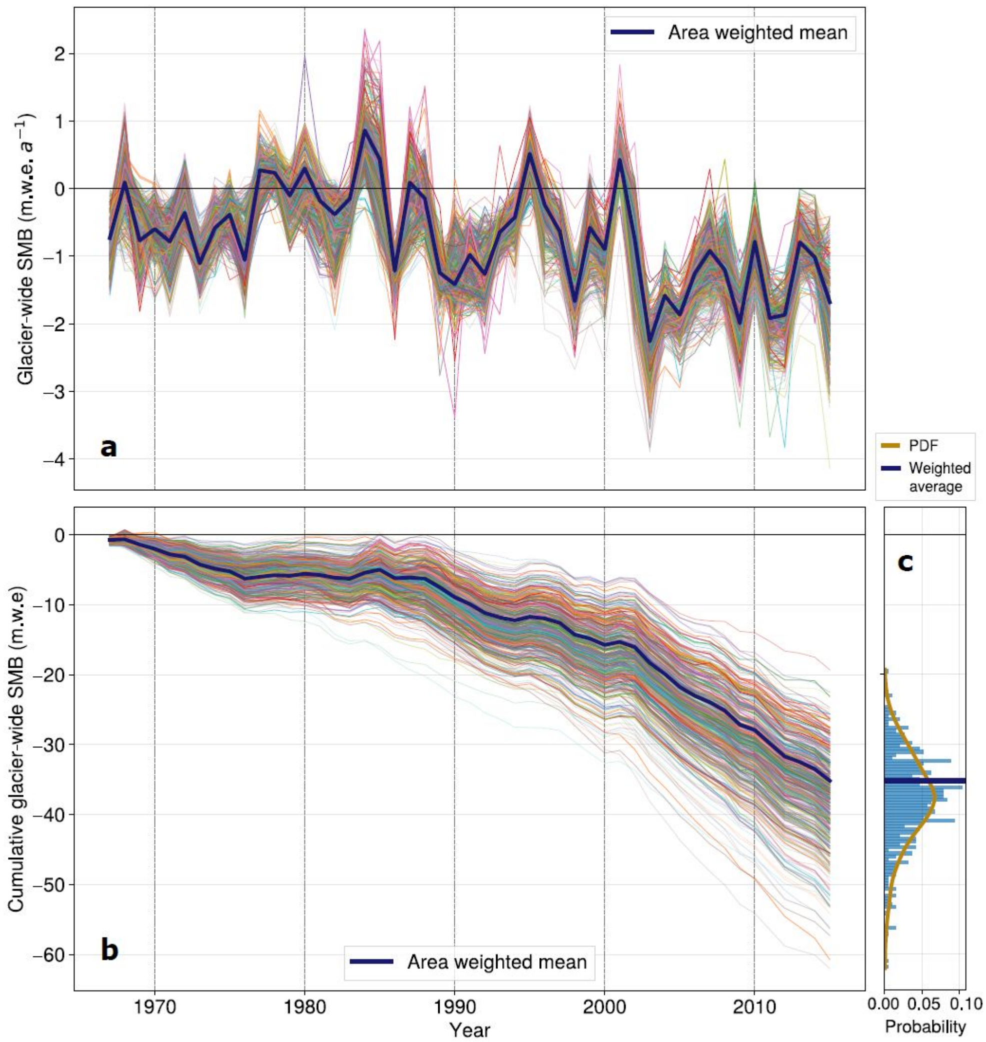


Fig. 3: why are no uncertainties included for the decadal averages?

Fig. 3 has been updated with decadal uncertainties.

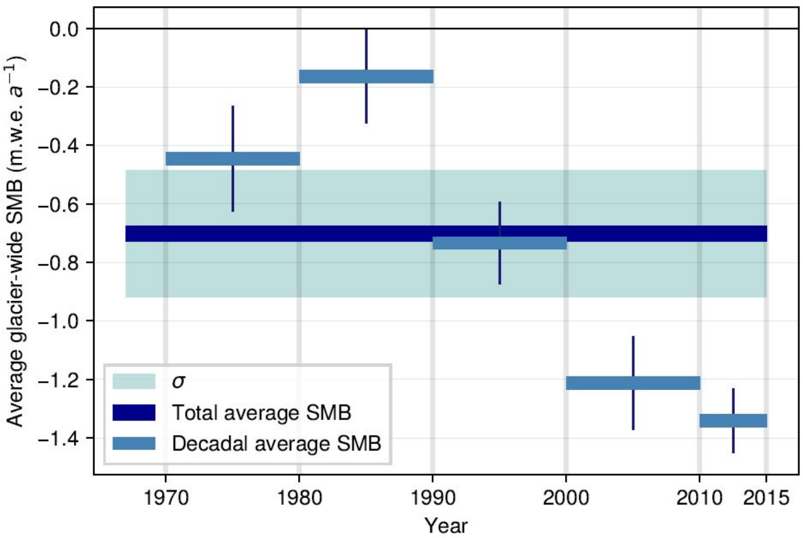
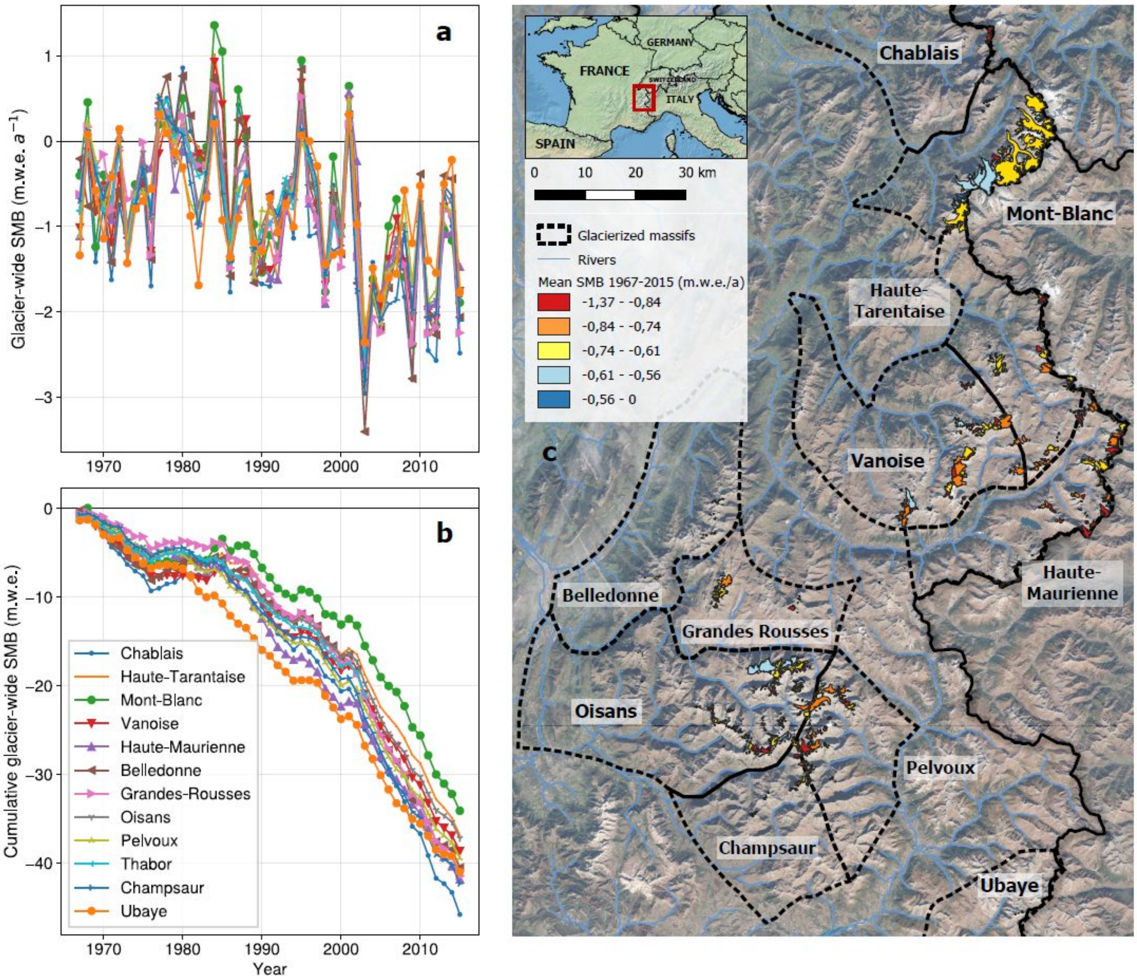


Fig. 4: great figure! But a bit busy (just visually); would it be possible to mute the background image a bit (and then perhaps change the text color to black) so that the colors of the glaciers stand out more?

We do agree that Fig. 4 could be a little bit overwhelming. We have updated it following the suggestions of the reviewer. Now the colours of the glaciers are more visible, and it has a more homogenous feeling with all contour lines in black.



P8 L16: it’s more than three parameters: one local (the temperature sensitivity) and four global ones (precipitation correction factor, precipitation lapse rate, temperature threshold for solid precipitation, and melt temperature threshold); see Figs. 4-7 in Marzeion et al. (2012).

The sentence has been updated with the correct information as it follows:

“This model was optimized based on five parameters: the temperature sensitivity of the glacier (local); and a precipitation correction factor, precipitation lapse rate, temperature threshold for solid precipitation and melt temperature threshold (global)”

P8 L22: perhaps clarify that the 38 glaciers are not the global sample used for calibration.

The new section in the supplementary material has been updated following this suggestion:

“**M15U calibrated their model with global SMB observations, including 38 glaciers in the European Alps, most of them located in Switzerland for the 1901-2013 period**; in this study we used observations of 32 glaciers, all located in the French Alps for the 1967-2015 period.”

P8 L31: I believe that the CV results in the Marzeion et al. (2012) study are also influenced by the global “optimization” (see above) of the four parameters; probably, a focus on the Alps would have led to a different parameter choice, and hence different CV results.

Indeed, that is what we tried to convey with the warning to the readers. We are comparing a global model with a regional model, so the specificity of the calibration is completely different, giving a clear advantage to the regional model. We hope that with the updated sentence from the previous comment this will be more clear to the reader.

P10 L1 and following: another reason for the different behaviour around the 2003 “break point” might be that the Marzeion et al. (2012) model, by construction, cannot capture the lasting effect that the extreme 2003 year may have had on albedo; while your model may be able to capture this (I guess – I’m not sure) by essentially taking the time as an additional predictor?

Our mass balance model does not have any perception of time, as no time stamps are used as predictors. I believe the main reason(s), as stated in the article, are the fact that we use higher resolution climate forcings, which better capture the climate signal on the glaciers, and most importantly, that the deep learning SMB model is nonlinear, which gives it a greater deal of flexibility to simulate this kind of transitions compared to linear models. This was already observed during the cross-validation analysis in Bolibar et al. (2020), where the linear model with Lasso, which behaves similarly to a temperature-index model, showed biases at the beginning and end of the 1984-2015 period, as the parameters were calibrated to fit the whole period, which presents rather neutral SMBs at the beginning and strongly negative SMBs by the end (see the Figure below taken from Bolibar et al., 2020). The nonlinear deep learning SMB model showed much lower biases, further demonstrating that the climate and glacier systems are highly nonlinear.

Authors reply to Ben Marzeion on "A deep learning reconstruction of mass balance series for all glaciers in the French Alps: 1967–2015"

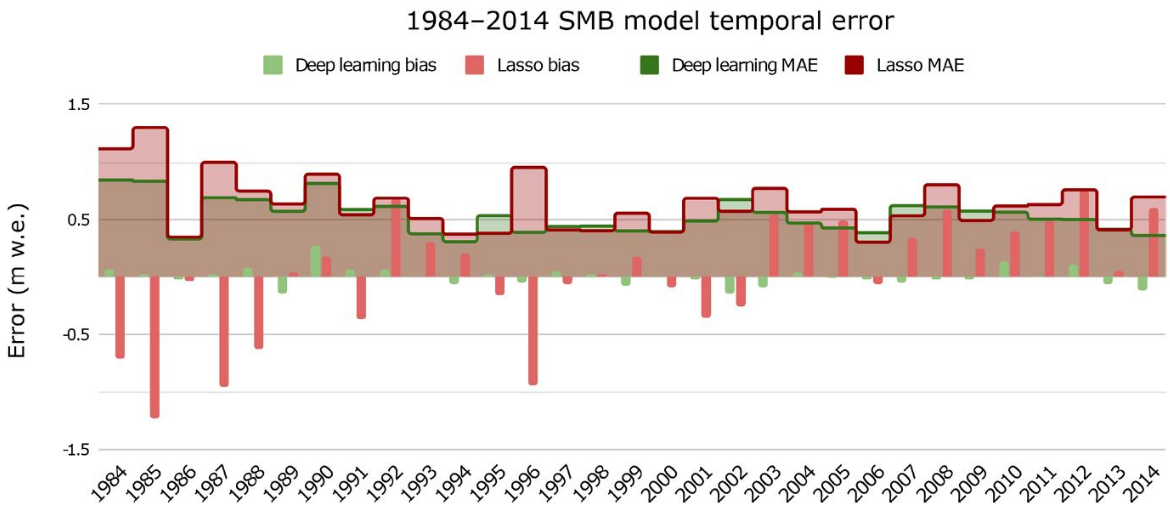


Fig. S2: would it be possible to re-arrange the legend such that it is easier to compare the "B" to the "M" lines (e.g., shift the lowest line in the legend to the right)?

The legend in Fig. S4 (previously S2) has been updated in order to have the "B" and "M" lines on different columns.