The most important thing about expedition food is that there is some.
Eric Shipton

Abstract

Fire in the hole

Acknowledgements

Fire in the hole

Contents

Chapter 1

Introduction

The most important thing about expedition food is that there is some.
Eric Shipton

Why are we concerned about glaciers?

"Changes in the longer-lived components of the cryosphere (e.g., glaciers) are the result of an integrated response to climate, and the cryosphere is often referred to as a 'natural thermometer'. But as our understanding of the complexity of this response has grown, it is increasingly clear that elements of the cryosphere should rather be considered as a 'natural climate-meter', responsive not only to temperature but also to other climate variables (e.g., precipitation). However, it remains the case that the conspicuous and widespread nature of changes in the cryosphere (in particular, sea ice, glaciers and ice sheets) means these changes are frequently used emblems of the impact of changing climate." This citation, quoted from the fifth Intergovernmental Panel on Climate Change (IPCC) assessment report (**?**), highlights the gap between the striking observations of massive retreat of glaciers and the in-depth understanding of the causes of these changes.

As a first step, good observations of glacier mass changes are needed to make a reliable diagnostic of the past and recent glacier mass changes. The glacier contribution to the current sea level rise (SLR) for the beginning of the twenty-first century was only recently comprehensively assessed (**?**), and the future evolution of glaciers and ice sheets is the main source of uncertainty in the SLR projections (**?**). For the beginning of the twenty-first century, and more precisely for the 2003-2009 period, glaciers contributed to 30 % of the observed SLR (**?**). Then, the second step is to develop models of glacier evolution, which can help to understand the processes responsible for glacier changes, and for example attribute the share of anthropic forcings (**?**). They can predict the future evolution of glaciers under different climate scenarios (**????**). Third, the glacier evolution models can be further used as inputs for hydrological models to assess the impacts of glacier changes on water resources for populations living downstream at a local scale (**???**).

However, the lack of good observational data disrupts the implementation of glacier

models, this is the case for High Mountain Asia (HMA) glaciers (**??**).

Why in High Mountain Asia?

The lack of glacier measurements, despite extensive glacier coverage, is all the more problematic, as HMA glaciers sustain the river discharge during the dry months for some densely populated basins (**????**), and therefore, realistic projections of HMA glacier changes are crucially needed. Satellite based techniques can partially alleviate the lack of field studies, but they are limited to pluri-annual averages (**??**).

Under a similar CO_2 emission scenario (Representative Concentration Pathway (RCP) 4.5), different glacier models predict a mass loss of 49 to 55 % of the current glacier mass for the entire HMA by 2100 (**????**). The good agreement among these models for the end of the century is surprising, because they are calibrated with different strategies, and they strongly differ for the early twenty-first century mass changes. For instance, **?** model prediction for 2000-2016 is more than twice as negative as **?** observation (for the period 2003-2009) on which **?** model is calibrated. This example rises important questions about the models calibration and about the relevance of the processes modeled.

The large tongues of HMA glaciers are often covered by a thick layer of debris (**?**), which effect was included in only one of the above mentioned models (**?**).

Why a focus on debris?

At the scale of HMA, the model prediction of **?** does not significantly differ from the other models, despite an explicit modeling of the debris effect on ice ablation. However, at the scale of the local Dudh Koshi catchment, two studies found irreconcilable results for the future of glaciers by 2100 (**??**). Different choices in glacier modeling led to the prediction of glacier mass reduction of 8-10 % in one case (**?**) and 84-95 % in the other case (**?**). The two studies are not directly comparable, because they investigated different areas and used different climate change inputs, but the main source of discrepancy is the modeling of the debris effect on ablation and the modeling of debris transport in one case (**?**).

Glaciological knowledge is based mostly on debris free glaciers, but the extent of the debris cover is expected to increase in a context of global warming, with a widespread slowdown of glacier tongues (**?**), which favors debris emergence (**????**). The recent increase in debris cover extent has been documented, for example, in the Alps (e.g., **??**), in Garhwal (e.g., **?**) and in the Everest region (e.g., **?**).

Consequently, within the course of the coming years, we might partially change our vision of glacier tongues, and debris. It was before considered as anecdotal feature, but has become very common. The potential influence of debris on the glacier evolution is still unclear, and it is therefore needed to better understand the relationship between debris and glacier mass balance. Within this long-term prospects, the aim of this PhD work is to assess the recent evolution of HMA glaciers and to quantify the influence

of debris on the glacier mass balance. This work is based on a multi-scale approach where large scale observations help to build a statistical intuition and/or validate models behavior, while, in parallel, fine scale approaches are developed to study processes, even if they are localized and their conclusions are not easy to extrapolate.

A short note to the reader

This manuscript organization follows a general direction from large scales to small scales. It starts with a review of the current state of the art knowledge about HMA climate and glaciers (chapter **??**), at the end of which the detailed research questions addressed in this manuscript can be found. The main body consists in three chapters, each of which is based on an article (one published, one accepted and one in review). The articles are introduced by a short note and for some of them I present further development and research directions. A conclusion summarizes this work and provides future research directions (chapter **??**). An article published in 2016 and based on a work I did for my master's thesis is appended at the end of the manuscript, as it was an important basis for the chapter **??**.

This structure implies some repetitions among the different chapters and a couple of inconsistencies, such as the use of m w.e. a^{-1} or m w.e. yr^{-1} for the mass balance units, which are imposed by the different journal styles (both in compliance with **?**).

Chapter 2

Deep learning applied to glacier evolution modelling

All models are wrong, but some are useful.
George Box

Preface

2.1 Abstract

We present a novel approach to simulate and reconstruct annual glacier-wide surface mass balance (SMB) 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.

2.2 Introduction

Glaciers are arguably one of the most important icons of climate change, being climate proxies which can depict the evolution of climate for the global audience (**?**). 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 (**???**). 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 (**?**). 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. (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. **?**), 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. **??**).
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. **?**).

(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. **?**) to more complex ones based on different approaches, such as a calibrated equilibrium-line altitude (ELA) model (e.g. **?**), a glacier retreat parameterization specific for glacier size groups (**?**) or volume/length-area scaling (e.g. **??**).
2. Process-based models, like GloGEMflow (e.g. **?**) and OGGM (e.g. **?**), 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. **?**), approach glacier dynamics by explicitly simulating physical processes and solving the full Stokes equations (e.g. **??**).

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., **?**). 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 (**?**). 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 (**??**). 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., **??**), climatology (e.g., **??**) and hydrology (e.g., **??**), we believe that the glaciological community has not yet exploited the full capabilities of these approaches. Despite this fact, a number of studies have taken steps towards statistical approaches. **?** 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 **?** and **?** for the simulation of glacier length instead of mass balances. Later on, **?** 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 (**?**). 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 **?**. These deeper ANNs, which remain unexploited in glaciology, allow to capture more nonlinear structures in the data even for relatively small datasets (**??**).

Here, we present a parameterized regional open-source glacier model: the ALpine Parameterized Glacier Model (ALPGM, **?**). When most glacier evolution models tend to incorporate more and more physical processes in SMB or ice dynamics (e.g., **??**), 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 h methodology (e.g., **?**) is used, for which we dynamically compute glacier-specific h 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, **?**) while the initial ice thickness distribution of glaciers are taken from **?**, 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.

2.3 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.

2.3.1 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 models generated by the Glacier-wide SMB simulation component.

2.1 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 climate predictors from the stored files, retrieves the topographical predictors from the mul-

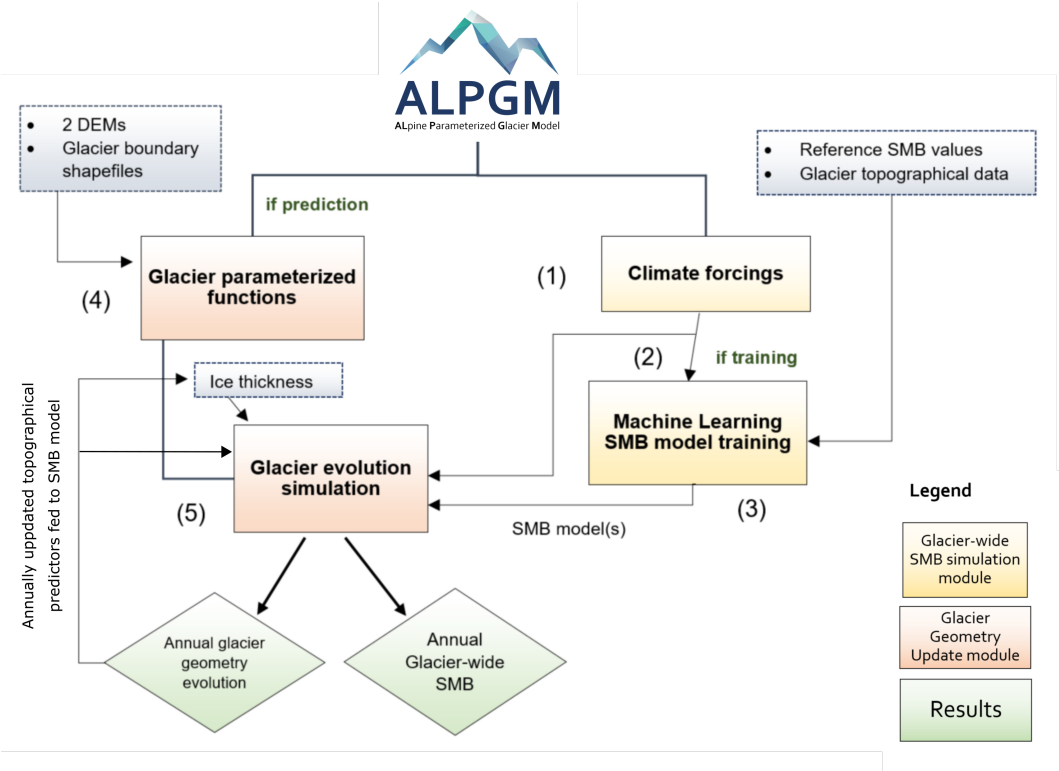

Figure 2.1: ALPGM structure and workflow

titemporal 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 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-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.

2.3.2 Glacier-wide surface mass balance simulation

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.

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, 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, 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 (**?**). Such combinations can be expressed following Eq. (1):

$$SMB_{g,y} = f(\hat{\Omega}, \hat{C}) + \varepsilon_{g,y} \tag{2.1}$$

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, $SMB_{g,y}$.

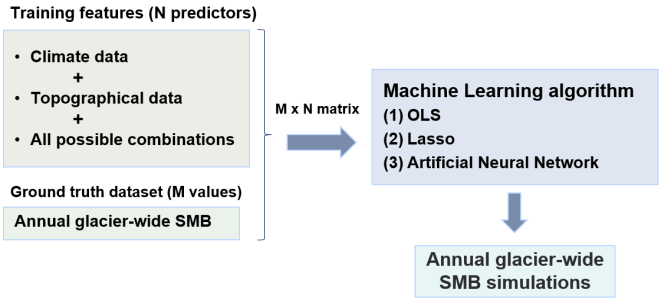

Figure 2.2: Glacier-wide SMB simulation component workflow. Machine learning models are dynamically created based on training data

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) (**?**); and (3) a deep Artificial Neural Network (ANN). ALPGM uses some of the most popular machine learning Python libraries: StatsModels (**?**), Scikit-learn (**?**) and Keras (**?**) with a TensorFlow backend. The overall workflow of the machine learning glacier-wide SMB model production in ALPGM is summarized in Fig. 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 (**?**) 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 (**?**). 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 (**?**). These criteria ensure that the models explain as much variability as possible, avoid collinearity and are statistically significant.

Lasso

The Lasso (Least absolute shrinkage and selection operator) (**?**) 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 (**?**). Because of its properties, it strikes a balance between subset selection (like all-possible regressions) and Ridge regression (**?**). 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 (**?**), can also be chosen with a classical cross-validation.

Deep artificial neural network

Artificial neural networks (ANNs) are nonlinear statistical models inspired by biological neural networks (**??**). 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 (usually nonlinear) activation functions (**?**). 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 **?**, **?**, as well as **??** 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 the training deep neural networks, that lead to better representation of complex data patterns. 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. Their ability to model complex functions of the input parameters makes them particularly suitable for modelling complex nonlinear systems such as the climate system (**?**) and glacier systems (**?**).

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 (**?**). 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

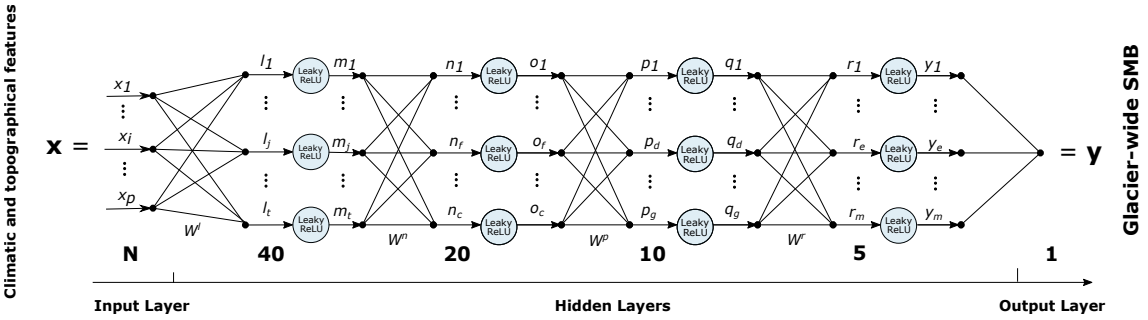

Figure 2.3: Deep Artificial Neural Network architecture used in ALPGM. The numbers indicate the number of neurons in each layer

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 (**?**), 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 reliability and the fact they help prevent the "dead ReLU" problem, where certain neurons can stop "learning" (**?**). The He uniform initialization (**?**) 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 (**?**).

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 (**?**). 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.

2.3.3 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 h parameterization. The idea was first developed by **?** and then adapted and implemented by **?**. 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 h 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 **?**, 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. **?** 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 (**????**), 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.

2.4 Case study: French alpine glaciers

2.4.1 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).

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 (**?**). 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 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 **?**. 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).

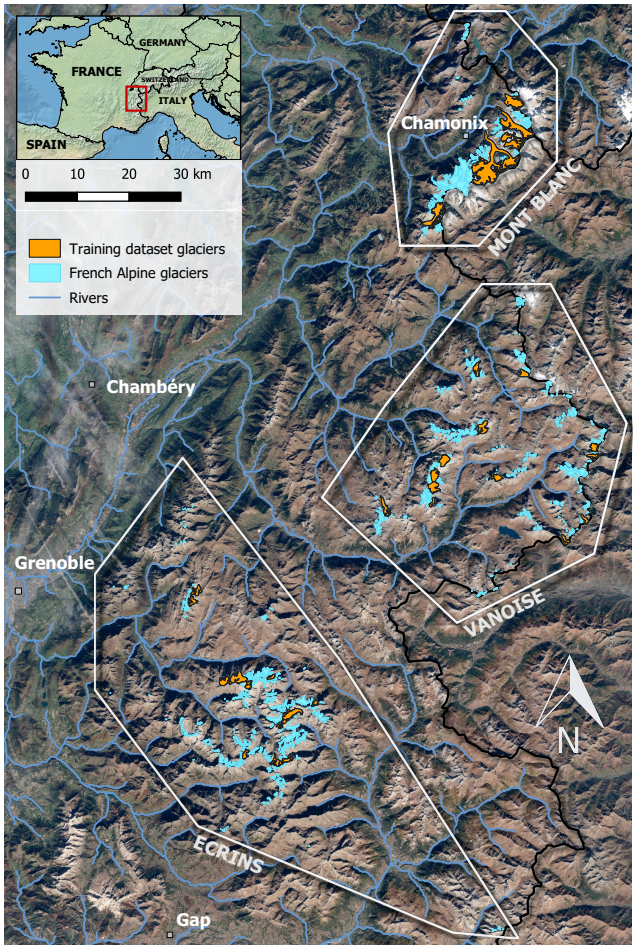

Figure 2.4: French alpine glaciers used for model training and validation and their classification into 3 clusters/regions (Écrins, Vanoise, Mont-Blanc). Coordinates of bottom left map corner: 44°32'N, 5°40'E, coordinates of the top right map corner: 46°08'N, 7°17'E.

Topographical glacier data and altimetry

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., **?**) partly available through the GLIMS Glacier Database (**?**). We worked with the 1967, 1985, 2003 and 2015 inventories (**?**, 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 (**?** 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 (**?**), and their uncertainties are negligible compared to many other parameters in this study.

Glacier ice thickness

Glacier ice thickness data come from **?**, hereafter F19, based on the Randolph Glacier Inventory v6.0 (RGI, **?**). 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, **?**) 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 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. 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 **?**. 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.

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 (**?**) 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).

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

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.

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 (**?**). **?** 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. **?** 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. **?** 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. **??** 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 **?**, 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. **?** 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 **?**, 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 $\hat{\Omega}$ and $\hat{C}$ from (Eq. 1) read as:

$$\hat{\Omega} = \left[\begin{array}{ccccccc} \overline{Z} & Z_{\max} & \alpha_{20\%} & \text{Area} & \text{Lat} & \text{Lon} & \Phi \end{array} \right] \tag{2.2}$$

$$\widehat{C} = \left[\begin{array}{ccccc} \Delta CPDD & \Delta WS & \Delta SS & \Delta \overline{T}_{\text{mon}} & \Delta \overline{S}_{\text{mon}} \end{array} \right] \tag{2.3}$$

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 = 0°)
$\Delta CPDD$: 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 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 $\Delta \overline{T}_{\text{mon}}$ and $\Delta \overline{S}_{\text{mon}}$.

$$\begin{aligned} SMB_{g,y} = {} & ((a_1\overline{Z} + a_2 Z_{\max} + a_3\alpha_{20\%} + a_4 Area + a_5 Lat + a_6 Lon + a_7\Phi + a_8)\Delta CPDD + \\ & (b_1\overline{Z} + b_2 Z_{\max} + b_3\alpha_{20\%} + b_4 Area + b_5 Lat + b_6 Lon + b_7\Phi + b_8)\Delta SS + \\ & (c_1\overline{Z} + c_2 Z_{\max} + c_3\alpha_{20\%} + c_4 Area + c_5 Lat + c_6 Lon + c_7\Phi + c_8)\Delta WS + \\ & d_1\overline{Z} + d_2 Z_{\max} + d_3\alpha_{20\%} + d_4 Area + d_5 Lat + d_6 Lon + d_7\Phi + d_8 + d_n\Delta \overline{T}_{\text{mon}} + d_m\Delta \overline{S}_{\text{mon}} + \varepsilon)_{\text{g,y}} \end{aligned} \tag{2.4}$$

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 $\Delta \overline{T}_{\text{mon}}$ and $\Delta \overline{S}_{\text{mon}}$ 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.

Causal analysis

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 shoulder season months (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, **?**). 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 (**?**). 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 (**?**). 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.

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. **?** already proved that a nonlinear ANN improved the results with respect a classic stepwise multiple linear regression. Here, we draw a similar comparison using more advanced methods for a larger dataset: OLS and Lasso as linear

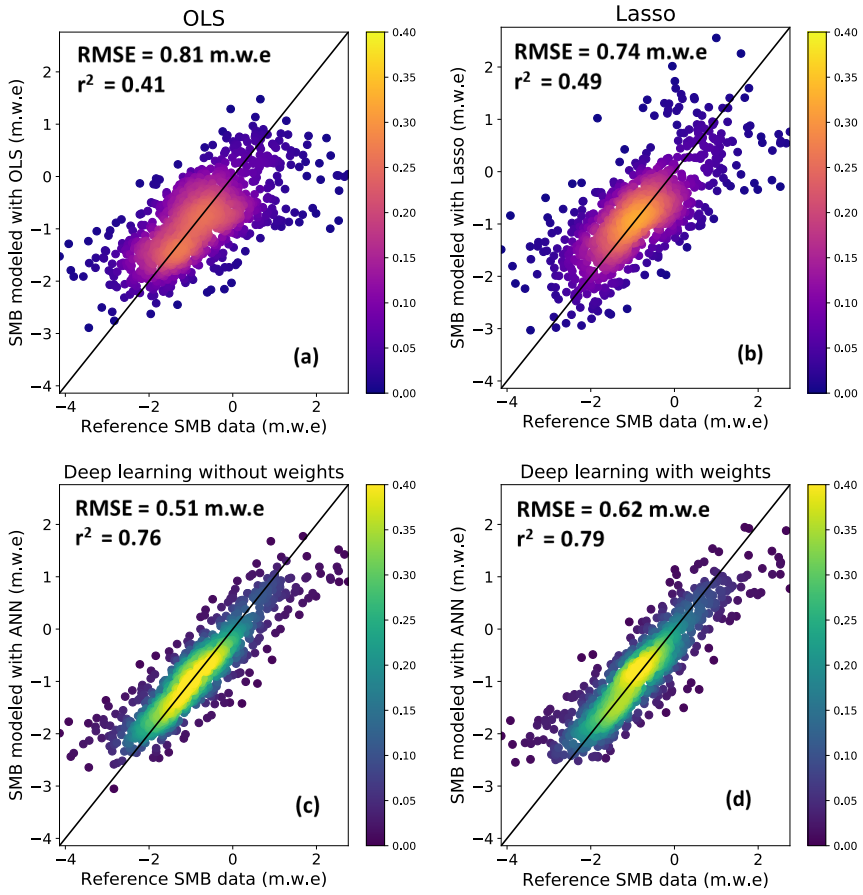

Figure 2.6: Evaluation of modelled annual glacier-wide SMB against the ground truth SMB data (both in m.w.e. a^{-1}) using Leave-One-Glacier-Out cross-validation. The colour (purple-orange for linear; blue-green for nonlinear) indicates frequency based on the probability density function. The black line 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)

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).

As a consequence, the added value of deep learning is especially relevant on glaciers

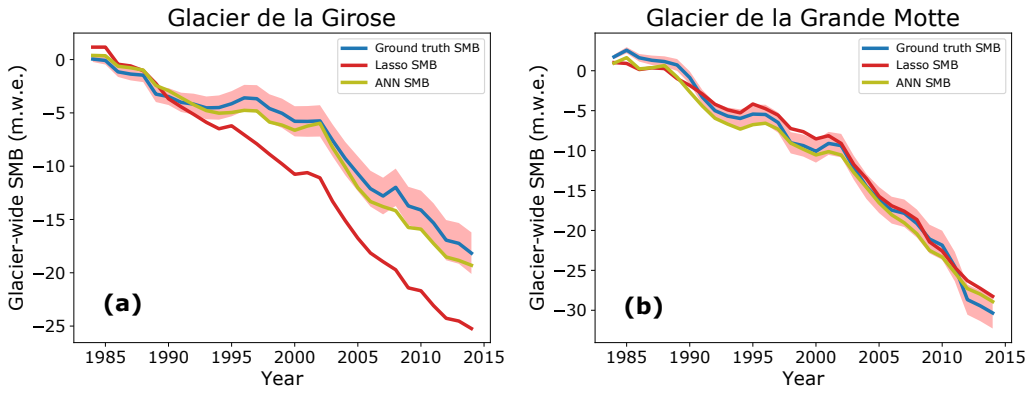

Figure 2.7: Examples of cumulative glacier-wide SMB (m.w.e.) simulations against the ground truth SMB data. The pink envelope indicates the accumulated uncertainties from the ground truth data. The deep learning SMB model has not been trained with sample weights in these illustrations.

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.

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 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 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 a high

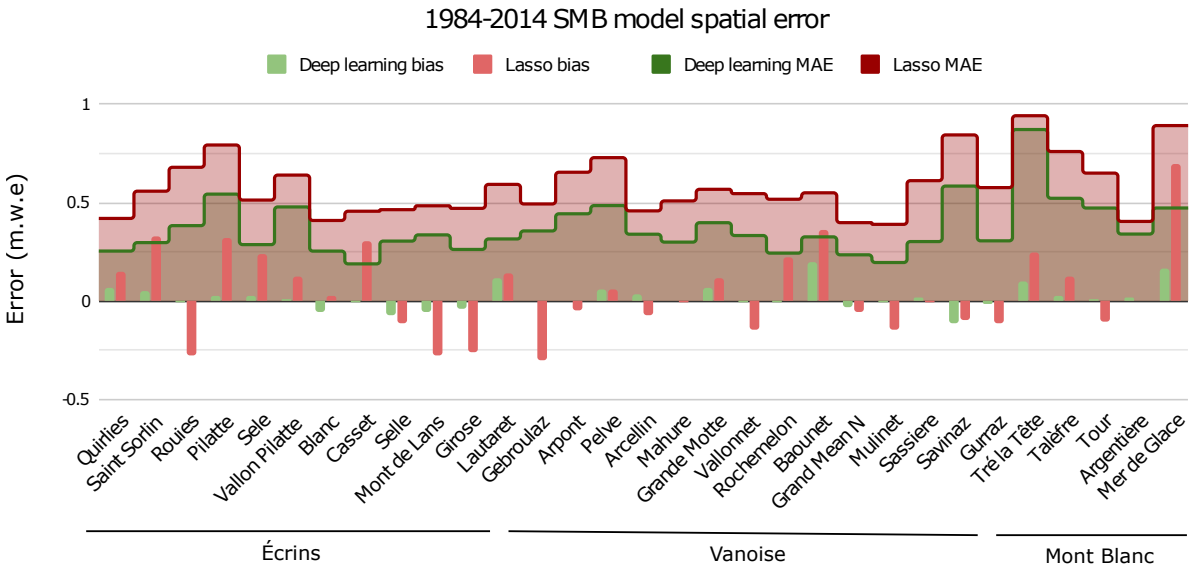

Figure 2.8: Mean average error (MAE) and bias (vertical bars) for each glacier of the training dataset structured by clusters for the 1984–2014 LOGO glacier-wide SMB simulation. No clear regional error patterns arise

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.*, years 1996, 2012 or 2014). 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.

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-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 (**?**).

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.

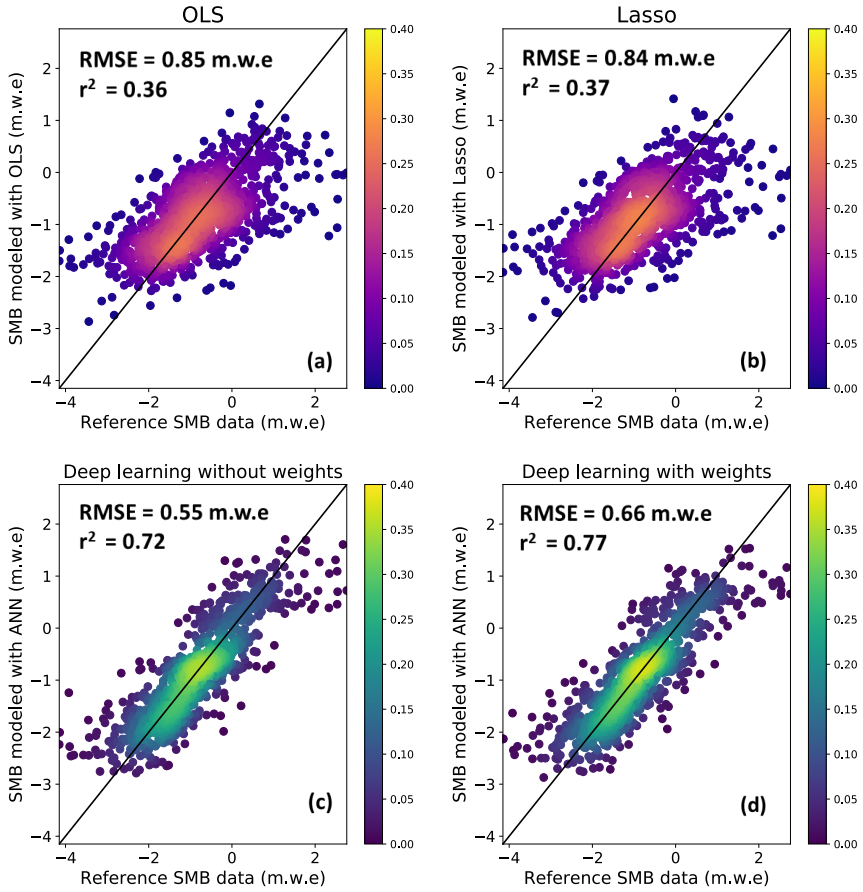

Figure 2.9: Evaluation of modelled annual glacier-wide SMB against the ground truth SMB data (both in m.w.e. a^{-1}) using Leave-One-Year-Out cross-validation. The colour (purple-orange for linear; blue-green for nonlinear) indicates frequency based on the probability density function. The black line 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).

2.4.3 Glacier geometry evolution: Validation and results

As mentioned in Sect. 2.3, the h parameterization has been widely used in many studies (e.g., **???????**). 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 h 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 **?** 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 h functions are computed for glaciers $\geq 0.5\ km^2$, which represented

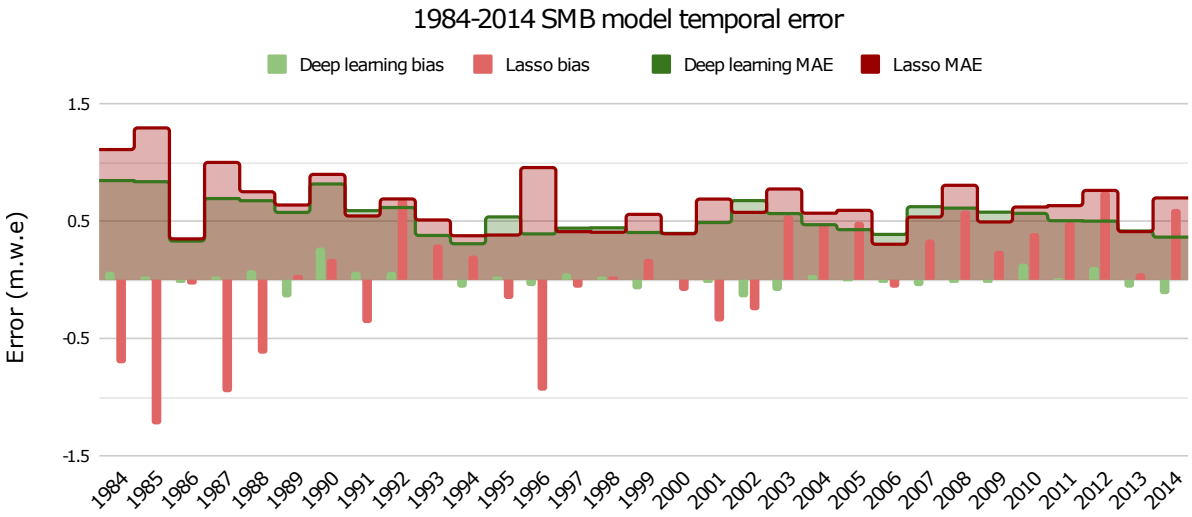

Figure 2.10: Mean average error (MAE) and bias (vertical bars) for each year of the training dataset for the 1984-2014 LOYO glacier-wide SMB simulation.

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 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 (**?**).

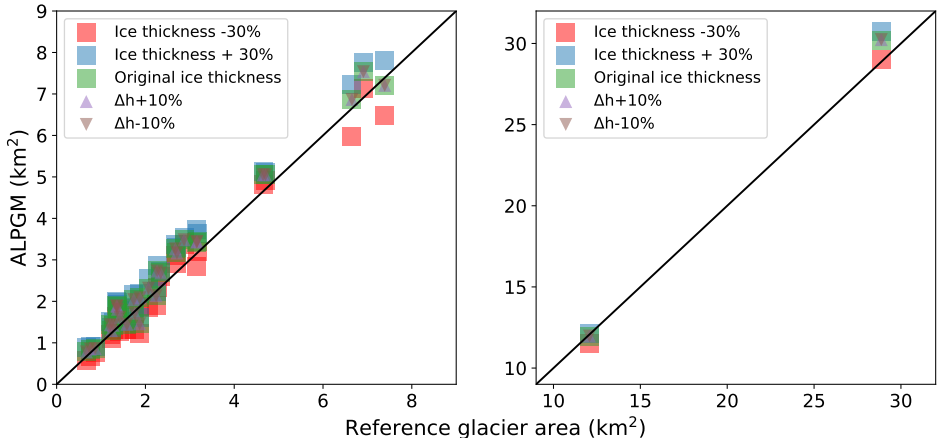

Figure 2.11: Simulated glacier areas for the 2003–2015 period for the 32 study glaciers using a deep learning SMB model without weights. Squares indicate the different F19 initial ice thicknesses used taking into account their uncertainties and triangles indicate the uncertainties linked to the glacier-specific geometry update functions. For better visualisation, the figure is split in two with the two largest French glaciers on the right.

2.5 Discussion and perspectives

2.5.1 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 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.

2.5.2 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 (**??**). 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 (**???**). 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.

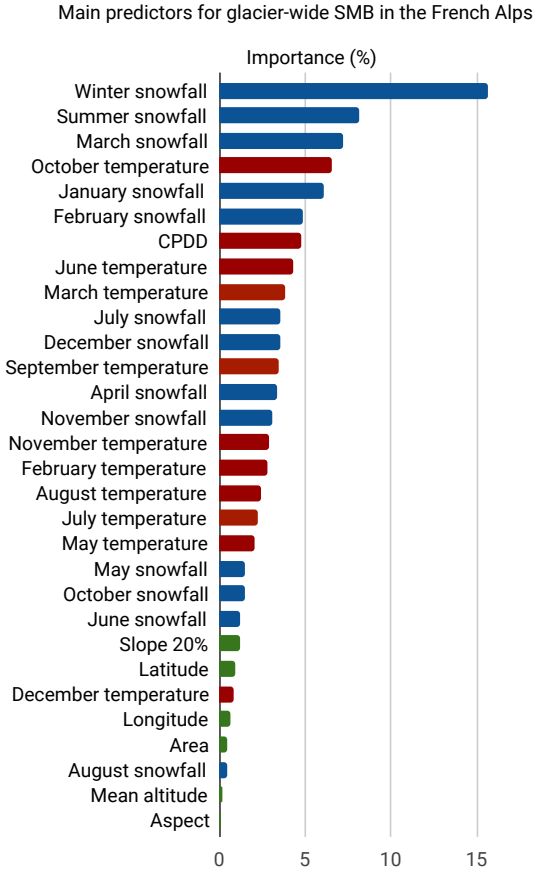

Figure 2.5: Contribution to the total variance of the 30 top topo-climatic predictors out of 55 predictors using Lasso. Green bars indicate predictors including topographical features, blue ones including accumulation-related features, and red ones including ablation-related features

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. 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. Con-

sequently, 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.

2.5.3 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., **?**), 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 (**?**). **?** 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.

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 m.w.e. $a^{-}1$ (against 0.51 from the full model). These results indicate that meteorological data is the primary source of information, determining the interannual high frequency variability of the glacier-wide SMB 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 and the topographical characteristics included in the glacier-wide SMB data (**?**). The fact that glacier-wide SMB is influenced by glacier topography poses the question of determining if the simulated glacier geometries can correctly reproduce topographical observations, needed to represent the topographical feedback present in glacier-wide SMB signals. These aspects are analyzed and discussed in Sect. 3 of the Supplementary material, showing small differences between the observed and simulated topographical parameters for the 2003-2015 period (Table S1). Additionally, the simulated glacier-wide SMBs using simulated topographical parameters show very small differences (0.069 m.w.e. $a^{-}1$ on average) compared to simulations using topographical observations (Fig. S6). Since glacier ice thickness esti-

mates date from the year 2003 (**?**), our validation period can only encompass 12 years. According to all the available data for validation, our model seems to be able to correctly reproduce the glacier geometry evolution, but since the 2003–2015 validation period is quite short, the validation performance might not be representative when dealing with future glacier evolution projections of several decades. Consequently, these aspects will have to be taken into account for future studies using this modelling approach for projections. Moreover, the cross-validation results of the SMB model(s) (Fig. 6–10) are representative of the performance of predictions using topographical observations. Despite the small differences found between simulated and observed topographical parameters, the SMB model's performance might be slightly different than the performance found in the cross-validation analysis. Therefore, it would be interesting for future studies to investigate the use of point SMB data, which could avoid the complexities related to the influence of glacier topography in glacier-wide SMB.

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. Even so, we believe that future efforts should be taken towards physics-informed data science glacier SMB and evolution modelling. Adding physical constraints in ANNs, with the use of physics-based loss functions and/or architectures (e.g., **?**), would allow to improve our understanding and confidence in predictions, reduce our dependency on big datasets, and to start bridging the gap between data science and physical methods (**????**). 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., **???**), 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 glaciarized regions could thereby be reconstructed, with potential applications in remote glaciarized regions such as the Andes or High Mountain Asia.

2.6 Conclusions

We presented a novel approach to simulate and reconstruct glacier-wide SMB series using deep learning for individual glaciers at a regional scale. This method has been included as a SMB component in ALPGM (**?**), 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.

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, upcoming methods adding physical knowledge to constrain neural networks (e.g., **??**) could provide interesting solutions to the limitations of our current method. By incorporating prior physical knowledge in neural networks, the dependency on big datasets would be reduced, and it would allow to transition towards more interpretable physics-informed data science models.

2.7 Supplementary material

2.8 Filtering of DEM rasters

Before computing the glacier-specific h parameterized functions, some preprocessing is done to the regional French Alps DEM raster files in order to filter artefacts and noise. The processing chain works as follows:

1. The regional DEM files are cropped using the 2003 glacier inventory shapefile outlines, thus obtaining glacier-specific rasters with the DEMs from 1979 and 2011.
2. The glacier surface altitude difference for this period (so-called dh/dt) which corresponds to the change in ice thickness is computed glacier by glacier by subtracting the two previously mentioned DEM rasters.
3. A first empirical filter is applied to all rasters to filter unrealistic values coming from artefacts (e.g., presence of clouds or saturation on the images used to generate de DEMs).
4. The filtered ice thickness difference (dh/dt) and DEM rasters are paired together as in Figure 12, and a low-order polynomial fit is applied in order to get the main trend of the scatterplot between the ice thickness difference vs. altitude.
5. A dynamic envelope/buffer around the polynomial fit line is computed for each

glacier based on a quantile between maximum and minimum values for each altitude. In order to smooth the computed envelope for each altitude, a convolutional filter is applied to these values in order to smooth them and to follow the polynomial fit. A dynamic sliding window size is used to adjust the averaging process to the characteristics of each glacier.

6. A second filter is then applied using the computed smoothed envelope buffer to remove outliers
7. A final polynomial fit is computed with a variable order depending on the number of remaining data values of each glacier.
8. The percentage of pixels of information available for computing the polynomial fit (the parameterized function) is displayed for each glacier at the end of the processing chain.

2.9 SMB statistical error analysis

In order to determine the error due to each predictor, a Lasso model was trained with the same training matrix as the ANN, but instead of using SMB as ground truth data the errors generated by the ANN without weights were used. As discussed in section 4.1, Lasso performs a regularization on the training dataset, thus keeping only certain predictors and removing the rest. By looking at the resulting coefficients of the model, we can estimate the linear contribution of each predictor to the final model error. Latitude and longitude appear as the most important error predictors, but their contribution might in fact indicate the different magnitude of errors between glaciers or regions, since the pair of coordinates specifically identifies each glacier. October, August and March temperature follow behind, indicating that changes in temperature during these months have an influence in the simulation errors. It is not surprising that two of these months appear as top predictors as seen in Fig. 5 as changes in temperature during these months at the transition between the accumulation and ablation season can have a strong importance on the surface mass balance processes.

Such an approach to analyze the influence of the different predictors into the quantification of uncertainties is of course limited, since a linear model is trained with nonlinear results. But these results are useful to determine the main contributors to errors rather than quantifying these errors, which has been done with the LOGO, LOYO and LSYGO cross-validations.

2.10 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:

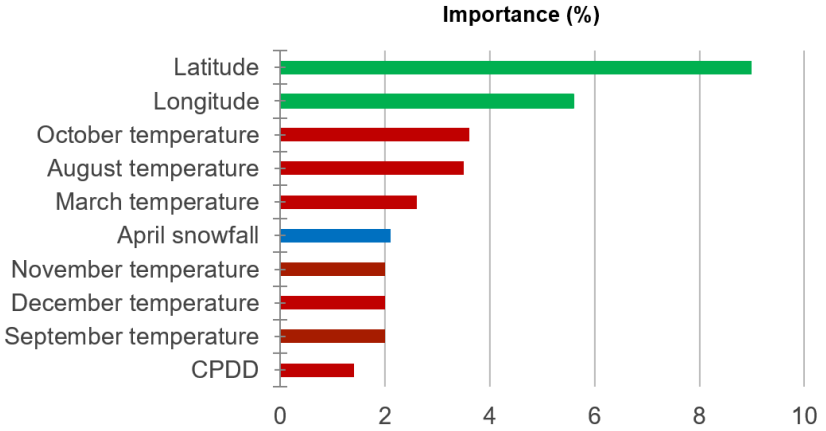

Figure 2.12: Importance (%) of the first 10 predictors using Lasso to predict residual error from the ANN SMB model. Green bars indicate topographical features, red bars temperature-related features and blue bars precipitation-related features.

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 MAE or mean difference 0.069 m.w.e a-1 1.8° 31.3 m 0.2 km2

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 km2 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.

2.11 Supplementary figures

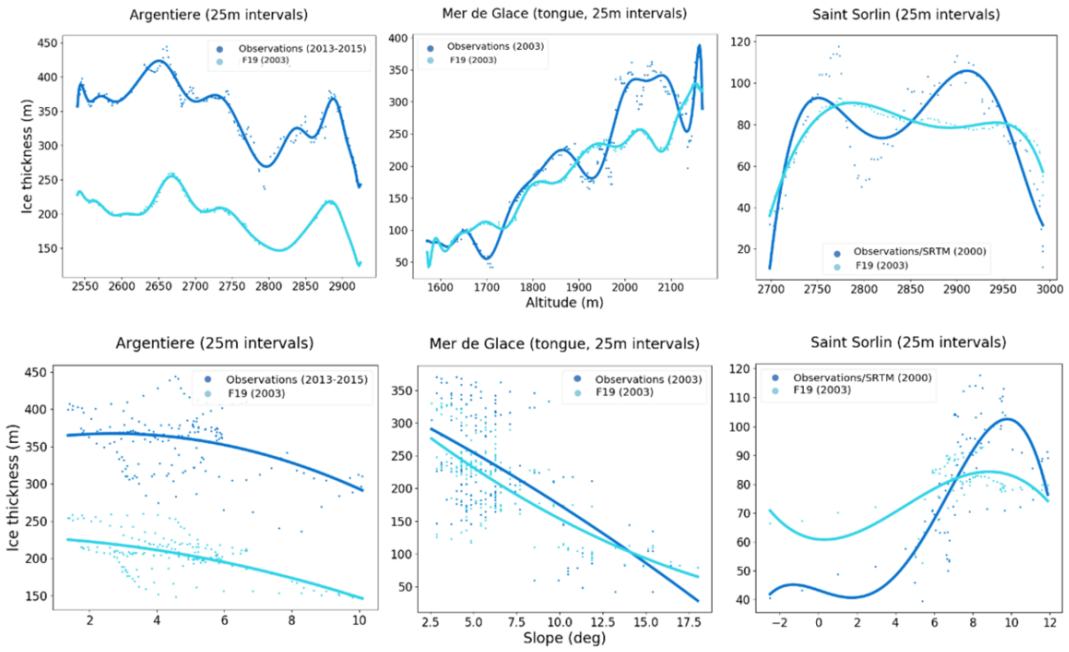

Figure 2.13: Comparison of simulated glacier ice thicknesses from F19 with observations from the GLACIOCLIM observatory. Points are compared at 25 m intervals on the glacier flowline. The polynomial fits have less degrees of freedom for the slope plots. Note that for some glaciers the dates are not the same

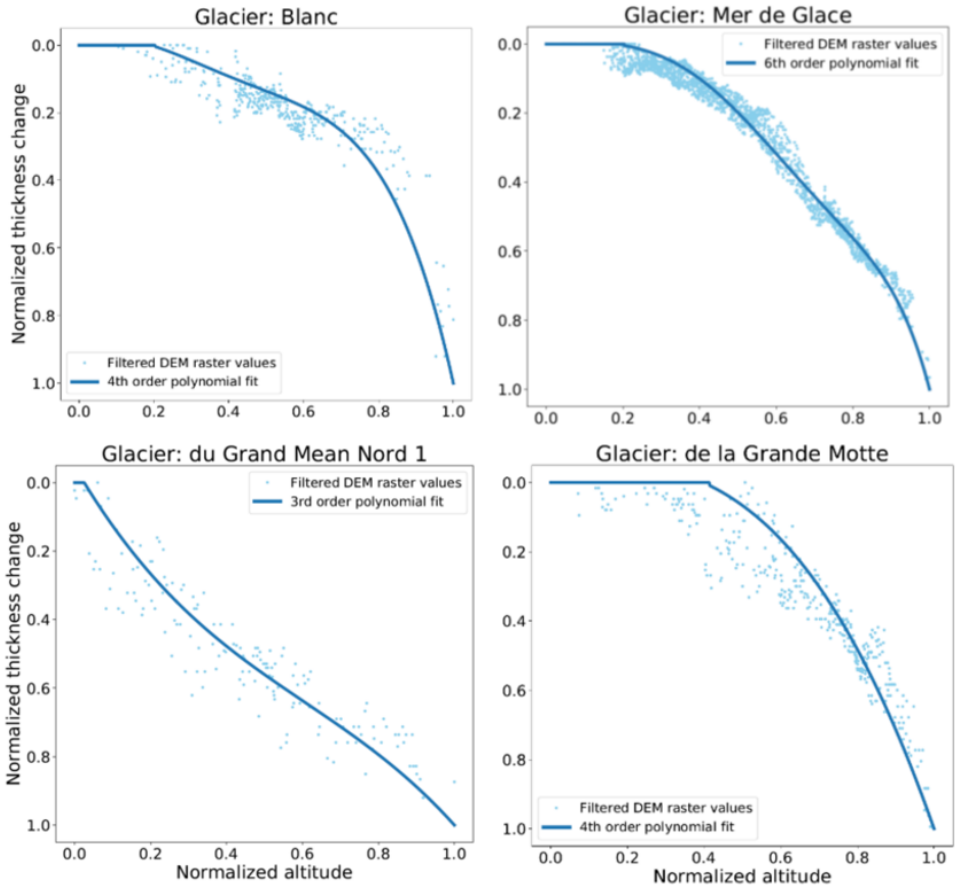

Figure 2.14: Examples of glacier specific h parameterized functions generated by ALPGM. The order of the polynomial fit depends on the number of available pixels.

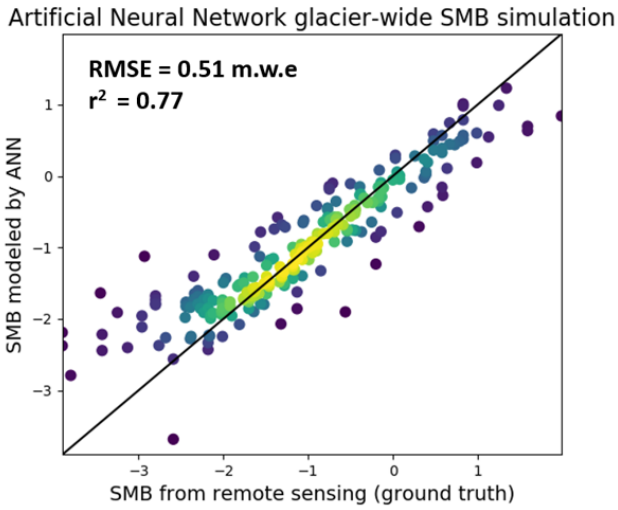

Figure 2.15: Results for the spatiotemporal cross-validation using Leave-Some-Glaciers-and-Years-Out (LSYGO). SMB values are in m.w.e. Compared to the other scatter plots from 3.2, there are less values available for test due to the severity of the spatiotemporal independence.

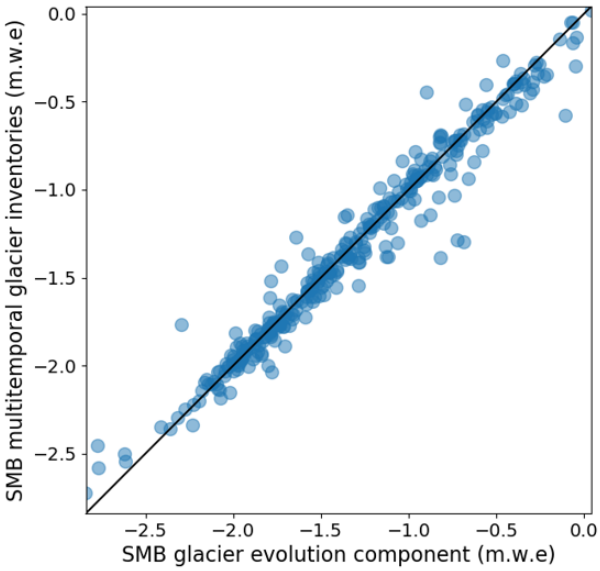

Figure 2.16: Comparison of glacier-wide SMB simulations (2003-2015, 32 case study glaciers) using topographical predictors from the multitemporal glacier inventories (Y axis) vs. using the full glacier evolution simulations in ALPGM with the Farinotti et al. (2019) ice thickness and DEM rasters (X axis). Average difference = 0.069 m.w.e. a-1

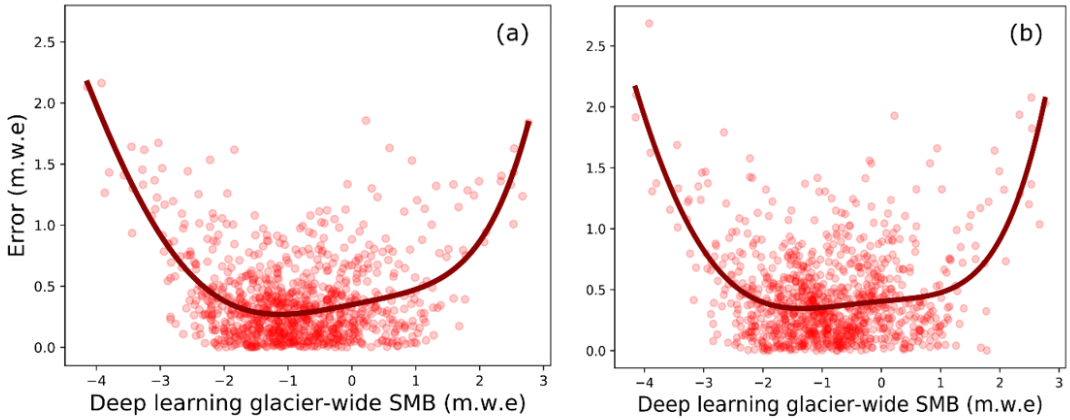

Figure 2.17: Error distribution of deep learning (without weights) glacier-wide SMB simulations for the 1984-2015 period for the 32 case study glaciers. (a) Performance in the spatial dimension using LOGO cross-validation; (b) performance in the temporal dimension using LOYO cross-validation. The red line corresponds to a 5th order polynomial fit.

Chapter 3

A deep learning reconstruction of mass balance series for all glaciers in the French Alps: 1967-2015

All models are wrong, but some are useful.
George Box

Preface

3.1 Abstract

Glacier mass balance (MB) data are crucial to understand and quantify the regional effects of climate on glaciers and the high-mountain water cycle, yet observations cover only a small fraction of glaciers in the world. We present a dataset of annual glacier-wide mass balance of all the glaciers in the French Alps for the 1967-2015 period. This dataset has been reconstructed using deep learning (i.e. a deep artificial neural network), based on direct MB observations and remote sensing annual estimates, meteorological reanalyses and topographical data from glacier inventories. The method's validity was assessed previously through an extensive cross-validation against a dataset of 32 glaciers , with an estimated average error (RMSE) of 0.55 m.w.e. a^{-1}, an explained variance (r^2) of 75% and an average bias of -0.021 m.w.e. a^{-1}. We estimate an average regional area-weighted glacier-wide MB of -0.69±0.21 (1σ) m.w.e. a^{-1} for the 1967-2015 period, with negative mass balances in the 1970s (-0.44 m.w.e. a^{-1}), moderately negative in the 1980s (-0.16 m.w.e. a^{-1}), and an increasing negative trend from the 1990s onwards, up to -1.26 m.w.e. a^{-1} in the 2010s. Following a topographical and regional analysis, we estimate that the massifs with the highest mass losses for the 1967-2015 period are the Chablais (-0.93 m.w.e. a^{-1}), Champsaur (-0.86 m.w.e. a^{-1}) and Haute-Maurienne and Ubaye ranges (-0.84 m.w.e. a^{-1} both), and the ones presenting the lowest mass losses are the Mont-Blanc (-0.68 m.w.e. a^{-1}), Oisans and Haute-Tarentaise ranges (-0.75 m.w.e. a^{-1} both). This dataset - available at: https://doi.org/10.5281/zenodo.3925378 (**?**) - provides relevant and timely data for studies in the fields of glaciology, hydrology

and ecology in the French Alps, in need of regional or glacier-specific annual net glacier mass changes in glacierized catchments.

3.2 Introduction

Among all the components of the Earth system, glaciers are some of the most visibly affected by climate change, with an overall worldwide shrinkage despite important differences between regions (Zemp et al., 2019). The European Alps are among the regions with the strongest glacier mass loss over recent decades, with expected mass losses between 60% and 95% by the end of the 21st century (**?**). These major glacier mass changes are likely to have an impact on water resources, society and alpine ecosystems (e.g. **???**). In order to study and quantify all these potential consequences, the availability of glacier mass balance data is of high relevance. Therefore, open historical datasets are crucial for the understanding of the driving processes and the calibration of models used for projections. Unlike glacier length, glacier mass balance (MB) provides a more direct indicator of the climate-glacier interactions (**?**). Glacier surface mass balance (SMB) is classically measured using the direct or glaciological method, by separately determining the ablation and accumulation totals. Direct measurements quantify the surface mass balance at different points of the glacier, and these values must be integrated at the glacier scale in order to assess the glacier-wide SMB (**?**). These different point SMB measurements can show a high nonlinear variability, which can complicate this integration process towards glacier-wide estimates (**?**). Moreover, field measurements require a lot of manpower, time and economic resources in order to be sustained for a meaningful period of time. On the other hand, recent advances in remote sensing allow estimating glacier MB changes at a regional level with unprecedented efficiency using geodetic and gravimetric methods (**?????**). Due to constraints related to the availability of digital elevation models (DEMs) or airborne data, these mass balance estimates normally encompass several years or decades. Some studies are bridging the gap towards an annual temporal resolution (**???**), but the coverage is still limited to glaciers without cloud cover or acquisition-related artefacts. This means that these mass balance datasets are often restricted to certain glaciers and years within a region. All these new datasets are extremely beneficial for data-driven approaches, fostering the training of machine learning models capable of capturing the regional characteristics and relationships (**?**). This type of approach allows to fill the spatiotemporal gaps in the MB datasets, therefore, it can be seen as a complement to remote sensing and direct observations.

On the other hand, MB reconstructions have already been carried out in the European Alps, providing a basis for comparison between different approaches (see **?** for a compilation). Two studies include reconstructions in the European Alps, including the French Alps, over a substantial period of the recent past: **??** reconstructed annual MB series of all glaciers in the Randolph Glacier Inventory for the last century. 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. **?** presented an approach to extrapolate SMB series of a limited number of glaciers to the mountain-range scale. By comparing multiple methods, he found the best results with a multiple linear regression based on 6 topographical parameters. From this relationship he reconstructed area-averaged SMB series of all the glaciers of the European Alps between 1900-2100 and analysed the trends for the different alpine nations and different glacier sizes.

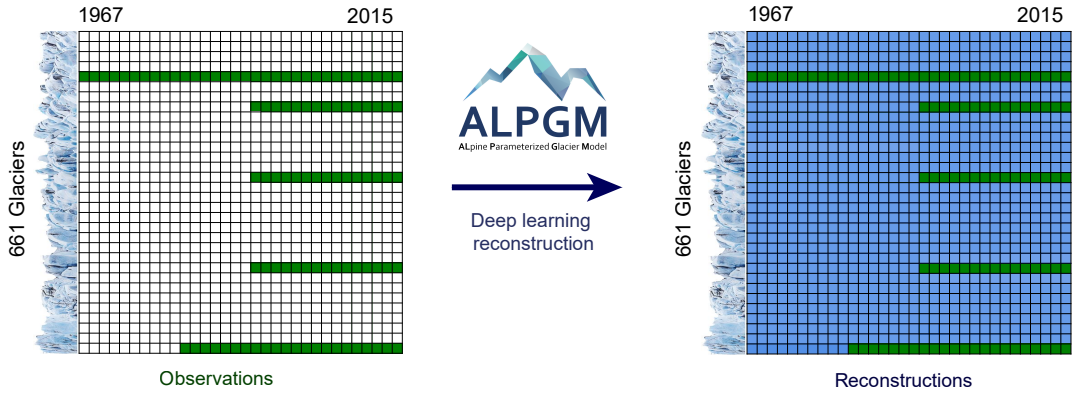

Figure 3.1: Summary of the deep learning regional MB reconstruction approach. From the available annual glacier-wide MB data, a deep learning model is used to reconstruct the full dataset, thus filling the spatiotemporal gaps in the observational dataset. Green indicates glaciers and years with MB observations and remote sensing estimates, and blue indicates reconstructed MB values. Glacier ice cliffs in the vertical axis indicate rows representing individual glaciers. The grid size with glaciers and years is schematic and only serves to illustrate the concept.

Here, we introduce a dataset of annual glacier-wide MB of all the glaciers in the French Alps (**?**), located in the westernmost part of the European Alps, between 5.08° and 7.67°E, and 44° and 46°13'N. Glacier-wide MBs have been reconstructed for the 1967-2015 period, using deep learning (i.e. a deep artificial neural network) (Fig. 1). This approach was introduced in **?**, for which a deep artificial neural network (ANN) was trained with data from 32 French alpine glaciers, as part of the ALpine Parametrized Glacier Model (ALPGM) (**?**). Annual glacier-wide MB values are reported for each glacier in the French Alps found in the 2003 glacier inventory (**?**). An overview of the methodology used to produce the dataset and a review of the associated uncertainties is presented in Sect. 2, followed by a dataset overview in Sect. 3, where the data structure and regional trends are described and where the dataset is compared to a previous study and observations.

3.3 Data and methods

3.3.1 Training data

For the reconstruction presented here, a dataset of 32 French alpine glaciers has been used for training, covering most of the massifs within the French Alps, which exhibit a great variability of topographical characteristics (Fig. S10). The French Alps are located in the westernmost part of the European Alps, rising from the Mediterranean sea northwards between 44 and 46°13' N, 5.08 and 7.67° E. Due to its particular geographical setup, glacierized mountain ranges in the French Alps have distinct climatic signatures. Southern glaciers exhibit a Mediterranean influence, whereas northern glaciers are mostly affected by western fluxes from the Atlantic, except for eastern glaciers close to the Italian border, which are more influenced by east returns.

Out of the 32 glaciers from this dataset, four glaciers include direct MB measurements from the GLACIOCLIM observatory, some of which since 1949. These direct observations have been calibrated using photogrammetric geodetic MB (**?**). On the other hand, 28 glaciers include estimates of annual glacier-wide MB from remote sensing between 1984 and 2014 (**?**). These remote sensing estimates were computed using (1) the end-of-summer snowline for every year, which in the European Alps is a proxy of the equilibrium-line altitude (ELA); and (2) geodetic MB for the 1984-2014 period quantified from two high-resolution DEMs. Both data sources are used to reconstruct the annual glacier-wide MB of each individual glacier for the same period of the geodetic MB.

This dataset of 32 glaciers, with a total of 1048 annual glacier-wide MB values, is used as a reference. Unlike point MB, glacier-wide MB is influenced by both climate and glacier geometry, producing complex interactions between climate and glacier morphology that need to be taken into account in the model. For each annual glacier-wide MB value available, the following data are compiled to train the ANN with an annual time step: (1) climate data from the SAFRAN meteorological reanalyses (**?**), 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 6 °C/km lapse rate. The updated temperature is then used to update the rain-snow parts from the same 300 m altitudinal band. Snowfall is considered as all precipitation fallen at temperatures equal or lower than 0° C. (2) annually interpolated topographical data between the 1967, 1985, 2003 and 2015 glacier inventories in the French Alps (update of **?**), with: mean and maximum glacier altitude, slope of the lowermost 20% altitudinal range of the glacier, surface area, latitude, longitude and aspect. Therefore, the topographical feedback of the shrinking glaciers is captured from these annually interpolated topographical predictors. These topoclimatic parameters were identified as relevant for glacier-wide MB modelling in the French Alps (**?**), and the dates of the glacier inventories determined the time interval for the reconstructions presented here.

For more details on the choice of predictors, the reader can find a more detailed analysis in **?**.

3.3.2 Methods

The annual glacier-wide MB dataset for the 661 French alpine glaciers has been reconstructed using a deep artificial neural network (ANN), also known as deep learning. ANNs are nonlinear statistical models inspired by biological neural networks (**??**). Recent developments in the field of machine learning and optimization enabled the use of deeper ANN architectures, which allows capturing more nonlinear and complex patterns in data even for small datasets (**?**). This modelling approach is part of the MB component of ALPGM (**?**), an open-source data-driven parameterized glacier evolution model. For a detailed explanation of the methodology, please refer to **?**. For the final reconstructions presented here, a cross-validation ensemble approach was used based on 60 Leave-Some-Years-and-Glaciers-Out (LSYGO) cross-validation models. Individual predictions of each of the members were averaged to produce a single out-

put. An ensemble approach has the advantage of further improving generalization, and reducing overfitting as well as the inter-model high variance typical from neural networks (**?**). A weighted bagging approach (**?**) was used in order to balance the dataset, giving more weight to under-represented data samples from the years 1967-1983. On the other hand, for the 32 glaciers with glacier-wide MB observations and remote sensing estimates used for training, an ensemble of 50 models trained with the full dataset was used, in order to achieve the best possible performance for this subset of glaciers, which represents a substantial fraction (45% in 2003) of the total glacierized surface area in the French Alps.

3.3.3 Uncertainty assessment

The uncertainties linked to the deep learning approach used in this study have been assessed through cross-validation, for which deep learning predictions were compared with observations and remote sensing estimates. A detailed presentation of the method's uncertainties and performance from the cross-validation study can be found in **?**. Block cross-validation ensured that all the 32 glaciers in the dataset were evaluated, with spatiotemporal structures formed by glaciers and years being considered in order to prevent the violation of the assumption of independence (**?**). This means that three different deep ANNs were produced: one for reconstructing glacier-wide MB in space, one for the reconstruction in time (future and past), and another one for both dimensions at the same time; each of these with a different calibration and performance. It was shown that the deep ANN performs better in the spatial dimension, in which the MB signal relationships with the predictors are the simplest. MB annual variability is mostly driven by climate, whereas geography and local topography (i.e. differences between glaciers) modulate the signal in space in a simpler way (**??**). Therefore, deep learning is capable of finding more structures in the spatial dimension, accounting for a better accuracy and explained variance compared to the temporal dimension. The deep ANN used in this study presents an RMSE of 0.55 m.w.e a^{-1} with an r^2 of 0.75 in LSYGO cross validation. The ANN MB reconstructions accurately reproduce the annual variability of glaciological observations from the GLACIOCLIM observatory (Figure S1). This reinforces the trust in the produced model ensemble, indicating that models trained with heterogeneous data comprised by glaciological and remote sensing estimates can correctly reproduce direct annual observations.

Nonetheless, only one glacier in the training dataset is smaller than 0.5 km^2 (Glacier de Sarennes, 0.3 km^2 in 2003), implying that uncertainties for very small glaciers (< 0.5 km^2) might differ from those estimated using cross-validation. In 2015, very small glaciers in the French Alps represented about 80% of the total glacier number, but they accounted for only 20% of the total glacierized area. This means that their importance is relative, for example in terms of water resources, but a user of this dataset should bear in mind that MB from these very small glaciers might carry greater uncertainties than the ones assessed during cross-validation. This might be especially true for extremely small glaciers (< 0.05 km^2) which can be considered as spatial outliers for the deep ANN. Since there is only one glacier with MB observations for very small glaciers and none for extremely small glaciers, there is no precise way to quantify these uncertainties. On the other hand, the ANN is mostly trained with glacier-wide MB data between 1984 and 2014, with a reduced amount of values between 1967 and 1984 (986

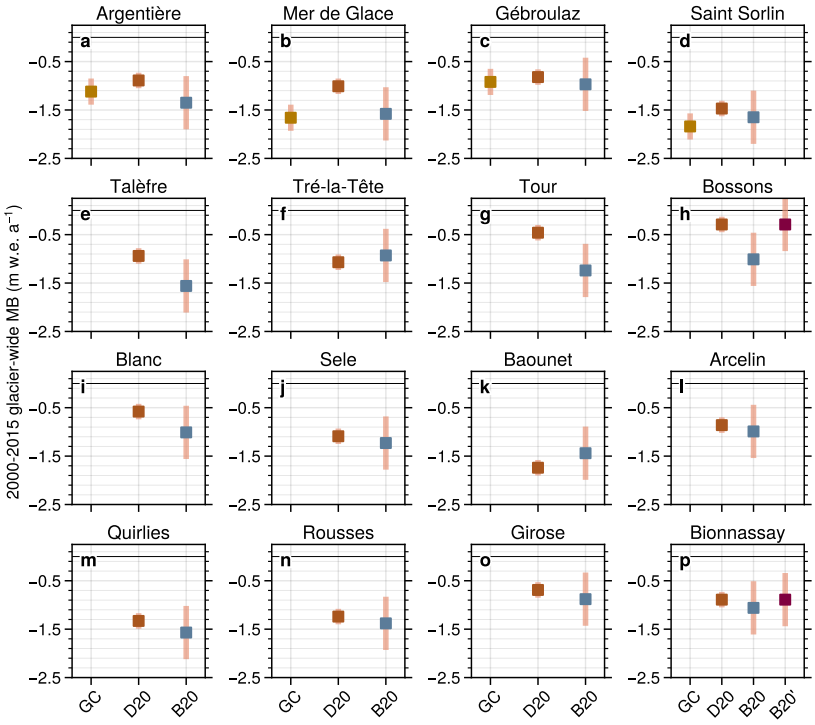

Figure 3.2: Comparison of average annual glacier-wide MB for the 2000-2015 period between the glaciological MB from the GLACIOCLIM observatory (GC), the ASTER-derived geodetic MB from Davaze et al., 2020 (D20), the MB reconstructions from this study (B20) and the reconstructions from this study recalibrated using the ASTER-derived geodetic MB (B20').

and 62 values, respectively). Since this early period contains on average more positive and neutral glacier-wide MB values than the 1984-2014 period, the performance of the ANN was specifically assessed for this period. An additional cross-validation was performed with four folds, each with a glacier including glacier-wide MB data before 1984. For each fold, all MB data of that glacier and time period were hidden from the ANN, and the simulated glacier-wide MBs between 1967 and 1983 were tested in order to assess the model's performance. The results showed that the ANN is capable of correctly reconstructing glacier-wide MB for glaciers and years before 1984 (Fig. S5), with an estimated accuracy (RMSE) of 0.47 m.w.e. a^{-1} and an estimated explained variance (r^2) of 0.65. This uncertainty assessment is based on roughly 10% of the full dataset, meaning that these estimates lack the robustness of the full cross-validation from **?**, but they serve to show that the model can accurately reconstruct glacier-wide MB data outside the main cluster of years used during training.

In order to further validate the reconstructions presented here, a comparison against independent ASTER (**?**) and Pléiades (**?**) geodetic MB data was performed, that helps to assess the bias of the MB reconstructions for the 2000-2015 (Fig. 2) and 2003-2012 (Fig. S2) sub-periods. The photogrammetric geodetic MB used to calibrate the MB datasets from **?** and the glaciological observations from GLACIOCLIM have a much higher resolution than ASTER-derived geodetic MB, but the comparison can bring interesting information for glaciers outside the training dataset. Our reconstructions show a good agreement with the geodetic MB for certain regions (e.g. Grandes Rousses),

except for some particular steep large high-altitude glaciers (e.g. Bossons and Taconnaz in the Mont-Blanc massif) that substantially differ from most glaciers in the French Alps. A more detailed analysis and additional figures comparing the MB datasets can be found in Sect. 1 of the Supplementary. In order to exploit this additional geodetic MB dataset, we have recalibrated our MB reconstructions for the 2000-2015 period using the ASTER-derived geodetic MB from **?** for some glaciers outside our training dataset (i.e. B20' in Fig. 2). Since ASTER-derived geodetic MB present important uncertainties for small glaciers (i.e. < 1 km^2), we have only recalibrated MB series for 16 large glaciers outside the training dataset with uncertainties lower than 0.15 m.w.e. a^{-1}. The calibration has been performed by adding the average annual bias between **?** and this study for the 2000-2015 sub-period.

3.4 Dataset overview

3.4.1 Dataset format and content

The MB dataset is presented in two different formats: (a) A single netCDF file containing the MB reconstructions, the glacier RGI and GLIMS IDs and the glacier names. This file contains all the necessary information to correctly interact with the data, including some metadata with the authorship and data units. (b) A dataset comprised of multiple CSV files, one for each of the 661 glaciers from the 2003 glacier inventory (Gardent et al., 2014), named with its GLIMS ID and RGI ID with the following format: *GLIMS-ID_RGI-ID_SMB.csv*. Both indexes are used since some glaciers that split into multiple subglaciers do not have an RGI ID. Split glaciers have the GLIMS ID of their "parent" glacier and an RGI ID equal to 0. Every file contains one column for the year number between 1967 and 2015 and another column for the annual glacier-wide MB time series. Glaciers with remote sensing-derived estimates (**?**) include this information as an additional column. This allows the user to choose the source of data, with remote sensing data having lower uncertainties (0.35±0.06 (σ) m.w.e. a^{-1} as estimated in **?**). Columns are separated by semicolon (;). All topographical data for the 661 glaciers can be found in the updated version of the 2003 glacier inventory included in the Supplementary material and in the dataset repository.

3.4.2 Overall trends

We estimate an average area-weighted regional glacier-wide MB of -0.69±0.21 (σ) m.w.e. a^{-1} between 1967 and 2015 (Fig. 3 and 4). As reported in previous studies (**???**), our reconstructed MB data show a slightly negative average value during the 1970s, even less negative in the 1980s, and then increasingly negative values in recent decades with an abrupt change in 2003 (Fig. 2). For this period (1967-2015), the year 2003 with its remarkable heatwave remains the most negative glacier-wide MB year (-2.26 m.w.e. a^{-1} on average), with 1984 being the most positive year of the study period (+0.85 m.w.e. a^{-1} on average). The area-weighted average MB is slightly less negative than the mean annual glacier-wide MB, showing a light asymmetry in the probability distribution function (PDF) (Fig. 3c).

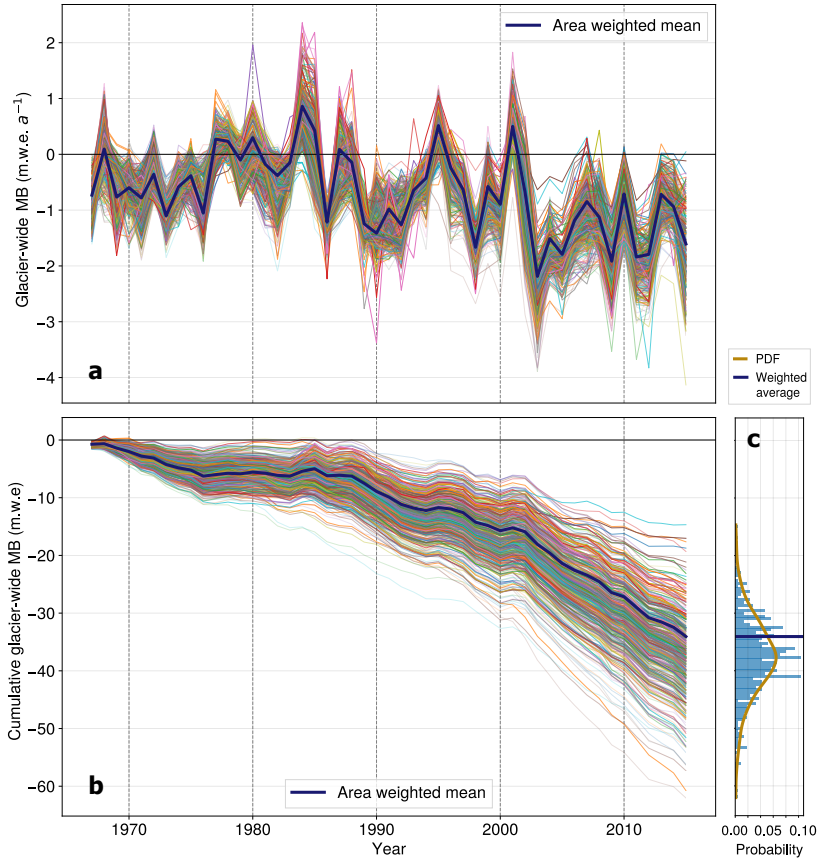

Figure 3.3: (a) Annual glacier-wide MB and (b) cumulative glacier-wide MB reconstructions of all the glaciers in the French Alps (N = 661) between 1967 and 2015. For each individual glacier, line thickness depends on glacier area, with smaller glaciers having thinner lines. The histogram (c) indicates the distribution and probability density function (PDF) of the 1967-2015 cumulative MB (m w.e.) of the dataset.

3.4.3 Regional and topographical trends

Here we analyse the main trends for the glacierized massifs and for some relevant topographical parameters. The reported glacier-wide MBs are only area-weighted if specifically mentioned. Interesting differences appear once the dataset is divided into mountain ranges (Fig. 5). The Mont-Blanc massif presents the lowest mass loss over the entire study period, with an average cumulative loss over the 1967-2015 period of 33.5 m.w.e. This is probably due to its northern location within the French Alps and its large high altitude accumulation areas, which resulted in more positive or less negative MBs, especially during the 1980-2000s. Oisans is the massif with the second lowest average cumulative mass loss (37.20 m.w.e.). Its glaciers have average altitudes ranging from 2290 to 3470 m.a.s.l., with around 50% of them having mean altitudes over 3000 m.a.s.l. and with about 40% of glaciers (including most of the large ones) having a northern aspect. Glaciers in Haute-Tarentaise present similar characteristics to those from Oisans, with mean altitudes ranging between 2300 and 3600 m.a.s.l., with about 60% of the glaciers above 3000 m.a.s.l. This less negative trend was especially important during the recent years with high mass losses from 2003 onwards. On the other hand, the Ubaye, Champsaur, Chablais and Haute-Maurienne massifs appear as

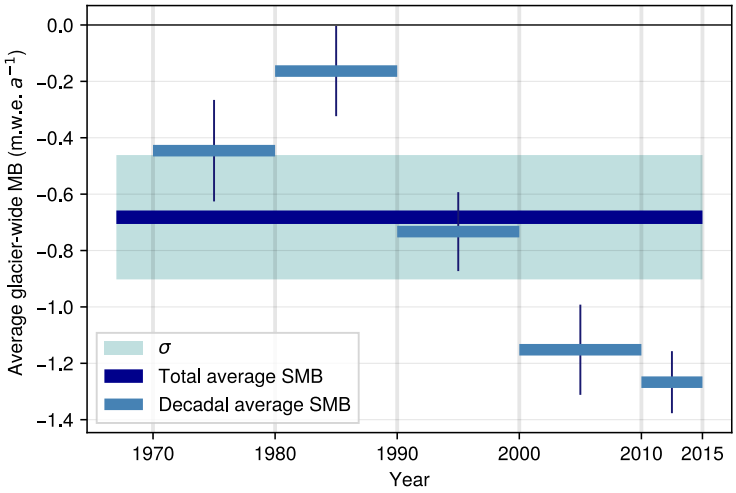

Figure 3.4: Averaged area-weighted decadal glacier-wide MB for the French Alps with decadal uncertainties. The total area-weighted glacier-wide MB is estimated for the 1967–2015 period.

the most affected mountain ranges with cumulative mass losses reaching between 41 and 46 m.w.e. for the four massifs over the 1967–2015 period. The Chablais range has a very small number of glaciers remaining, all of them at rather low altitudes (2200–2900 m.a.s.l.), relatively small (0.01 – 1.1 km^2), and with a northwestern aspect. Despite being the northernmost mountain range in the French Alps, its low altitude is most likely the main reason for the very negative MBs, which were under the regional average even during the positive years in the 1980s. The Champsaur range shows a similar situation, with very small glaciers (0.03 – 0.89 km^2) lying at relatively low altitudes (2300–3100 m.a.s.l.) in the southernmost latitudes of the Alps (44°7'). Finally, the situation of the Ubaye massif is quite similar to the one of Champsaur, being the southernmost glacierized massif in the French Alps, with a strong mediterranean influence. Such glaciers are remnants of the Little Ice Age, far from being in equilibrium with the warming climate, and can quickly lose a lot of mass through non-dynamic downwasting (**?**).

When classifying the MB time series by glacier surface area, we encounter the following patterns, with n being the number of glaciers in the subset and s its standard deviation: (1) Very small glaciers (< 0.5 km^2; n = 534; $\overline{MB}_{1967-2015}$ = -0.79 m.w.e. a^{-1}; s = 0.23 m.w.e. a^{-1}) present more negative glacier-wide MBs than (2) small/medium glaciers (ranging from 0.5 to 2 km^2; n = 93; $\overline{MB}_{1967-2015}$ = -0.74 m.w.e. a^{-1}; s = 0.18 m.w.e. a^{-1}) and (3) large glaciers (> 2 km^2; n = 34; $\overline{MB}_{1967-2015}$ = -0.68 m.w.e. a^{-1}; s = 0.14 m.w.e. a^{-1}) (Fig. S8). Very small glaciers present a larger spread of values than small/medium and large glaciers (s = 0.23 m.w.e. a^{-1} versus 0.18 and 0.14 m.w.e. a^{-1}, respectively). As explained in Sect. 2, the uncertainties for very small glaciers are greater due to their under-representation in the training dataset, meaning that analyses based on small glaciers have to be taken with greater care. The effects of these trends can be seen in the PDF of the cumulative MB reconstructions (Fig. 3c), where the area-weighted mean lies slightly outside the PDF maximum, showing how a great number of small glaciers are presenting higher losses. On the other hand, a clearer relationship between the glacier slope (computed here as the lowermost 20% altitudinal range slope) and glacier-wide MB arises, with steeper glaciers having less negative glacier-wide MBs

(Fig. S6 and S9). Glaciers with a gentle tongue slope generally present longer response times and higher ice thickness, which are associated with more negative mass balances (**???**). These results are in agreement with the findings by **?**, who computed the geodetic mass balance of all the Swiss glaciers for the 1980–2010 period. Overall, the topographical relationships found here are similar, although more negative than for the Swiss Alps (**??**), showing how the southernmost glaciers in the Écrins and Vanoise regions present stronger glacier mass losses. This is mostly due to their mediterranean climatic influence compared to the more continental Swiss and Austrian glaciers, which results in more negative MB in a warming climate (**?**). Nonetheless, results from this type of bivariate analysis can show rather biased trends, since the topographical variables are highly intercorrelated, with for example small glaciers having steeper slopes and *vice versa* (**?**). The position and evolution of the equilibrium line can totally reverse the trends of small or steep glaciers, so these relationships can strongly vary depending on the region or time period observed.

3.4.4 Comparison with previous studies and observations

In order to put into perspective the reconstructions presented in this study, we compare them to an updated version from the **?** reconstructions (B. Marzeion, personal communication, October 2019 - January 2020), and to all the available glacier-wide MB observations and remote sensing estimates in the French Alps. The goal of this comparison is not to draw conclusions on the quality of either reconstruction, but to analyse the differences among them and to try to understand the causes. In the updated version of **?** - referred as M_{15U} from now on - a global MB model relying on temperature and solid precipitation was used to reconstruct MB time series for all the glaciers in the world present in the Randolph Glacier Inventory (**?**). 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). As in **?**, the approach by M_{15U} was cross-validated respecting the spatiotemporal independence in order to evaluate its performance for unobserved glaciers and years. Due to the highly different methodologies and forcings of the two models, a direct comparison is not possible, so the following analysis is focused on the overall trends and sensitivities in the reconstructions and their potential sources. All the specific differences and details between the two models can be found in Sect. 2 from the Supplement.

The annual variability (Fig. 6), driven by climate, is quite similar between the two reconstructions. Conversely, important differences are found for different subperiods in the amplitude of the area-weighted mean glacier-wide MB series. These differences are the greatest in the 1970s, 1980s and 2010s, with similar average values for the 1990s and 2000s (Fig. 6 and S7). M_{15U} presents less negative and more positive glacier-wide MB values in the 1970s, but on the contrary, it presents more negative values in the 1980s compared to our results. We believe there might be two potential reasons for this: (1) In 1976 there was a shift in the winter mass balance regime in the French Alps, with more humid winters bringing more accumulation; and in 1982 there was a shift in the summer mass balance, resulting in increased ablation (**?**). Since both models use parameterized or statistical relationships for MB response to precipitation and temperature, they are likely to react differently to these changes. A similar situation is found

from the year 2003 onwards, where there was a substantial increase in temperatures and mass loss (e.g. **?**). Our reconstructions show a marked change in 2003 (change of slope in the cumulative plot in Fig. 6), whereas M_{15U} present a rather linear trend. The fact that M_{15U} used a volume-area scaling compared to the interpolated topographical data from inventories from this study means that the topographical feedback of the models might differ as well throughout the reconstructed period. (2) For the 1967–1983 interval, the amount of available glacier-wide MB data for training is much lower than for the rest of the period (green numbers in Fig. 6). This is likely the reason why the differences between our reconstructions and training data are greater for that period (Fig. 6). On the other hand, the similarities between our reconstructions and the training data for the 1984–2014 period are explained by the fact that the 32 glaciers with observations represent around 45% of the total glacierized area in the French Alps in the year 2003. For the periods before and after this interval, differences and uncertainties in the reconstructed values are greater because of the smaller sample size.

In the following, we argue that similarities between observations, remote sensing estimates and the reconstructed glacier-wide MB values for the 1984–2015 period in this study (Fig. 6) are not due to overfitting. First, for the vast majority of the 661 French glaciers, the reconstructions are based on an ensemble of cross-validated models, which intrinsically limits overfitting (see Sect. 2). Second, we analysed the deviation to the climatological mass-balance signal of the MB for each cluster of glacier-sizes. This analysis is presented in Sect. 3 of the supplementary material. It reveals that the similarities between the training data and the reconstructed glacier-wide MB values for the 1984–2015 period in Fig. 6 originate from big glaciers, that dominate both in the area-weighted reconstructions and in the training data (Fig. S3 and S4). However, for the other glacier-size classes, our reconstruction shows different patterns from the data in the training data, which suggests that the model is not overfitting (Fig. S3).

3.5 Conclusions

We presented a dataset of annual glacier-wide MB of all the glaciers in the French Alps (44° - 46°13'N, 5.08° - 7.67°E) for the 1967–2015 period (**?**). This dataset has been reconstructed using deep learning (i.e. an artificial neural network), based on direct and remote sensing annual glacier-wide MB observations and estimates, climate reanalysis and topographical data from multitemporal glacier inventories. The deep learning model is capable of reconstructing glacier-wide MB time series for unobserved glaciers in the same region based on patterns and structures learnt by the artificial neural network from the training data and their relationships with predictors. An extensive cross-validation was implemented to understand the characteristics of the MB signal in the region and to assess the method's validity and uncertainty. The average accuracy (RMSE) of the dataset is estimated at 0.55 m.w.e. a^{-1} with an explained variance (r^2) of 75%. Reconstructions show a mean area-weighted glacier-wide MB of -0.69±0.21 (1 σ) m.w.e. a^{-1} for the 1967–2015 period. Important differences are found among different massifs, with the Mont-Blanc (-0.68 m.w.e. a^{-1}), Oisans (-0.75 m.w.e. a^{-1} both) presenting the lowest mass losses and the Chablais (-0.93 m.w.e. a^{-1}), Champsaur (-0.86 m.w.e. a^{-1}) and Haute-Maurienne and Ubaye (-0.84 m.w.e. a^{-1} both) showing the highest losses. In order to put these results into perspective, this reconstruction was com-

pared to all available glacier-wide MB observations and remote sensing estimates in the French Alps as well as the physical/empirical reconstructions from another study (update from **?**). Interesting differences were found between the two methods, highlighting the different sensitivities and responses of different approaches to climate shifts that occurred during the study period. These differences are particularly relevant in the 1970s and 1980s, previous to a winter precipitation and summer temperature shift that occurred in the French Alps in the years 1976 and 1982, respectively. Moreover, after the famous 2003 European heatwave, glaciers experienced an acceleration in mass loss which is well captured by our reconstruction. This open glacier-wide MB dataset can be useful for hydrological or ecological studies in need of net glacier mass contributions of glacierized catchments in the French Alps. The publication of such open datasets is essential to future community-based data-driven scientific studies.

3.6 Supplementary material

3.7 Comparison with independent geodetic mass balance data

All available annual glacier-wide MB data in the French Alps have been used to train the MB ANN of the present study. However, some multi-annual geodetic mass balance (MB) datasets exist that can provide a means to validate the reconstruction's bias for specific glaciers during multi-annual time intervals. This type of analysis is more limited than the cross-validation done to annual glacier-wide MB values in **?**, as it only gives information about the bias of a sub-period of the reconstructions instead of the accuracy found via cross-validation. Our MB reconstructions are compared against ASTER geodetic MB from **?** for the 2000-2015 and 2003-2012 periods (Fig. 2 and S2) and against Pléiades geodetic MB from **?** for the 2003-2012 period (Fig. S2).

For certain glaciers, the ASTER and Pléiades geodetic MB give a less negative MB than the glaciological SMB used to train the deep learning SMB model. This fact might explain the slightly more negative trend of our reconstructions seen for the 2000-2015 and 2003-2012 periods, which experienced very negative MB after the well known summer 2003 heatwave. This is quite surprising, since both the GLACIOCLIM glaciological MB measurements and the annual glacier-wide MB data from Rabatel et al. (2016) have been calibrated with geodetic MB from photogrammetric DEMs, which have a very high spatial resolution. For some regions (i.e. Grandes Rousses), the independent geodetic MB are well within the uncertainty range of our model. However, large and steep glaciers in the Mont-Blanc massif and some other regions, such as Bossons, Talèfre and Tour display important differences. These glaciers have very large and high altitude accumulation areas, not seen in almost any glacier in our training dataset. On the other hand, several small glaciers present very important differences, with ASTER-derived MB being much less negative than our reconstructions. Data for small glaciers carry very large uncertainties, often of the same order of magnitude as the observations themselves. On top of that, flat or dome-type glaciers with large white areas with high reflectance present an important amount of noise, further increasing the associated uncertainty. This means that is quite hard to jump to conclusions from a direct comparison between

these glaciers and our reconstructions. The differences and influence of geodetic MB on the calibration of MB series should be properly studied, as they are often not taken into account as an additional uncertainty source. This topic goes beyond the scope of this study, but glacier modelling studies could benefit from integrating this in the list of uncertainties.

3.8 Model differences between the updated version of Marzeion et al. (2015) and this study

In order to contrast the results from Sect. 3.4, three important different aspects between our approach and the one of M_{15U} need to be highlighted:

1. M_{15U}'s model works with simplified physics, with a temperature-index model calibrated on observations; in this study we used a fully statistical approach based on deep learning, where physics-based considerations only appear in the predictor selection.

2. M_{15U} calibrated their model with global MB 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.

3. M_{15U} forced their updated model with CRU 6.0 (update of **?**), with 0.5° latitude/-longitude grid cells, which has a significantly lower spatial resolution and suitability to mountain areas than the SAFRAN reanalysis (**?**) used in this study, in which altitude bands and aspects are considered for each massif, and meteorological observations from high-altitude stations are assimilated.

The cross-validations of both studies determined a performance with an average RMSE of 0.66 m.w.e. a^{-1} and an r^2 of 0.43 for M_{15U} for the European Alps, and an average RMSE of 0.49 m.w.e. a^{-1} and an r^2 of 0.79 for this study. However, due to the highly different methodologies and forcings of the two models, a direct comparison is not possible, so the following analysis is focused on the overall trends and sensitivities in the reconstructions and their potential sources.

3.9 Influence of area in glacier-wide MB signal and proof on non overfitting

Due to similarities between the averaged reconstructed glacier-wide MB and the observations during the 1984–2015 period, we decided to include an analysis to isolate the topographical influence in the glacier-wide MB signal, in order to verify that the model is not overfitting. Since the climate signal is the main common driver of annual variability of glacier-wide MB in the region, one needs to find a way to isolate the topographical signal. In Fig. S3, the median reconstructed annual glacier-wide MB of the 661 glaciers in the French Alps (i.e. the annual variability, hence a proxy of the climate signal) is subtracted to the mean annual values of the observations and of 4 subsets of glaciers divided by area classes. Therefore, one can observe the residual influence

of glacier area on the glacier-wide MB signal. The influence of area on glaciers with observations is quite similar to glaciers with areas greater than 2 km^2, which is reasonable since glaciers with observations have an average of 4 km^2 (range: 0.3-31.8 km^2 in 2003). Moreover, one can see that even for a relatively short period of 30 years, the differences between the reconstructions for very small glaciers (< 0.5 km^2) and observations are quite important, accounting for an average cumulative loss of more than 5 m.w.e. As stated in Sect. 2, this does not necessarily mean that the model has fully captured the topographical influence in the glacier-wide MB signal in the region, but it does prove that the model is not overfitting since it exhibits consistent variations in MB when the topographical predictors move away from the training data. Moreover, this is coherent with the importance attributed to topographical predictors (**?**).

The same analysis has been performed with the reconstructions from the updated version of Marzeion et al. (2015) (Fig. S4). The gradient with respect to glacier surface area appears to be similar, except for the behaviour of glaciers after 2007. Small and middle sized glaciers (0.1 - 2 km^2) switch to a positive influence, as opposite to large glaciers (> 2 km^2), which transition to a negative influence. Conversely, our results show a more continuous trend, without a change of behaviour in the last years of the analysed period.

3.10 Supplementary figures

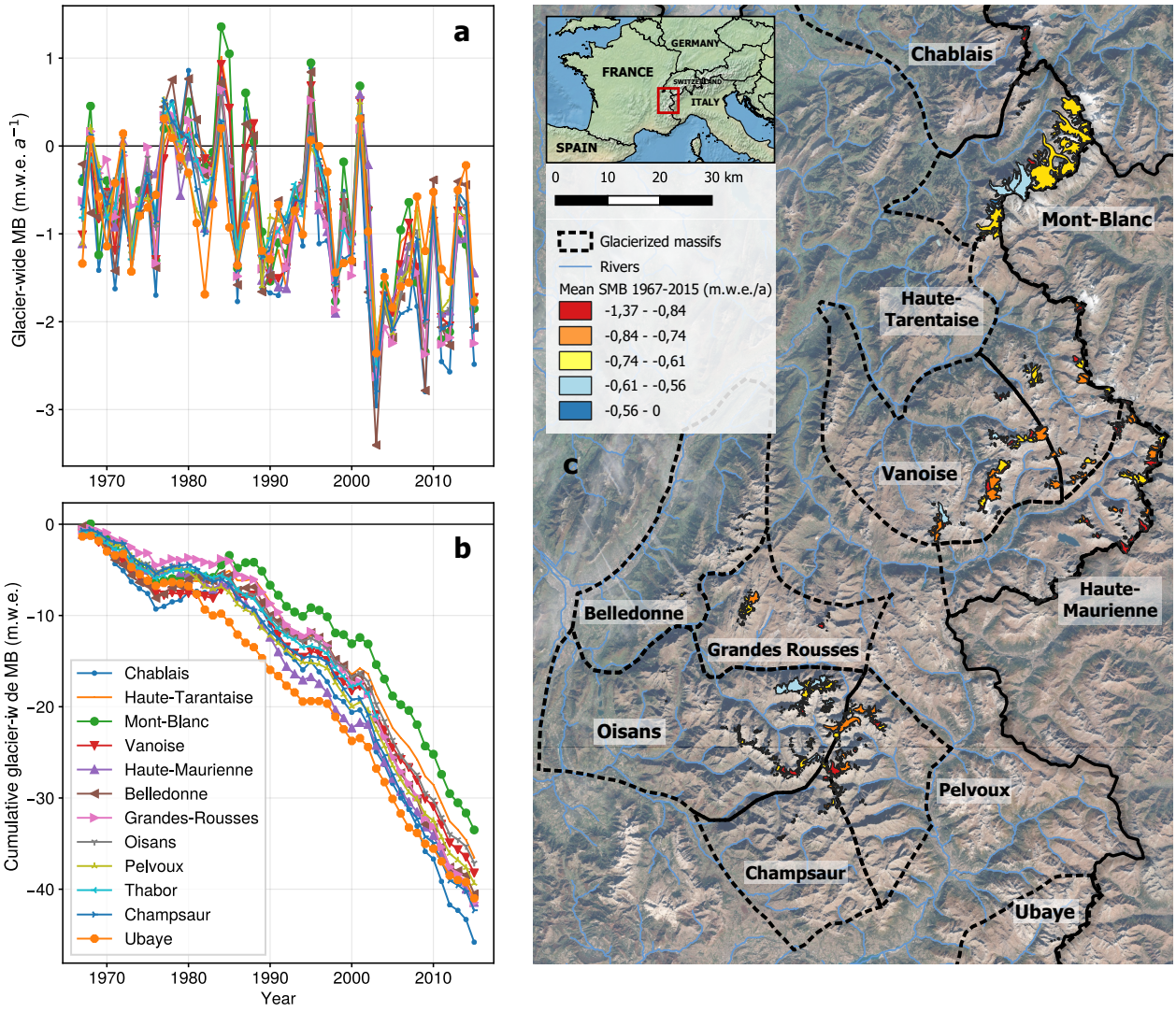

Figure 3.5: (a) Averaged annual glacier-wide MB and (b) cumulative averaged glacier-wide MB time series for each of the massifs in the French Alps between 1967 and 2015. (c) Glacierized massifs in the French Alps with the average glacier-wide MB for the 1967-2015 period. Coordinates of bottom left map corner: 44°32' N, 5°40' E. Coordinates of the top right map corner: 46°08' N, 7°17' E.

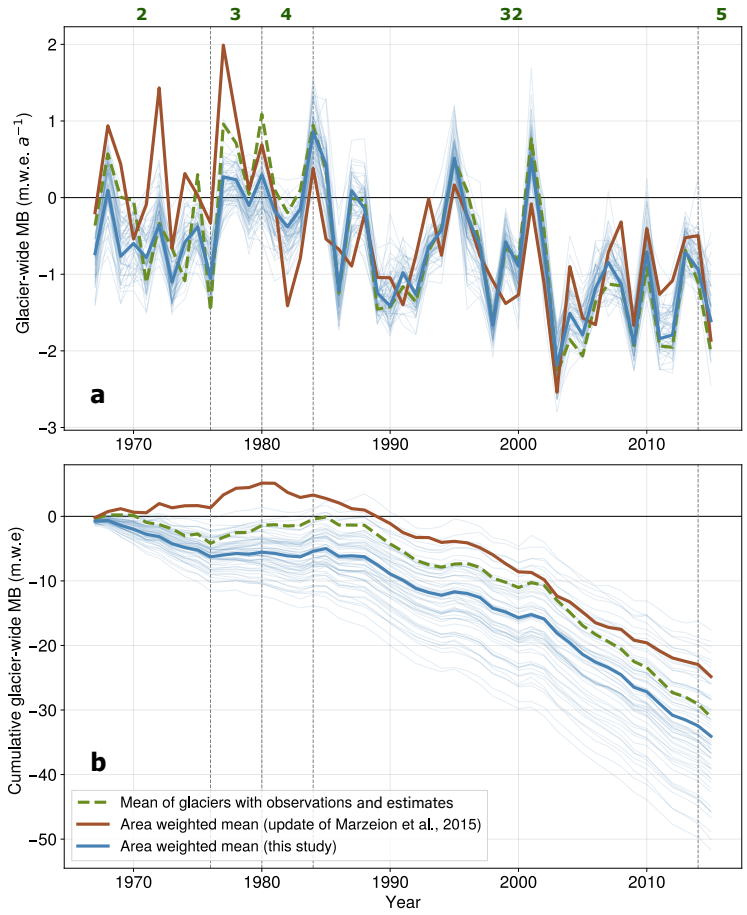

Figure 3.6: Comparison of (a) annual and (b) cumulative glacier-wide MB simulations in the French Alps between this study, reconstructions from an update from Marzeion et al. (2015) and the mean of all observations and remote sensing estimates available in the French Alps. Green numbers indicate the number of glaciers with MB observations and remote sensing estimates for each period and thin light blue lines indicate the area-weighted mean of each of the cross-validation ensemble members.

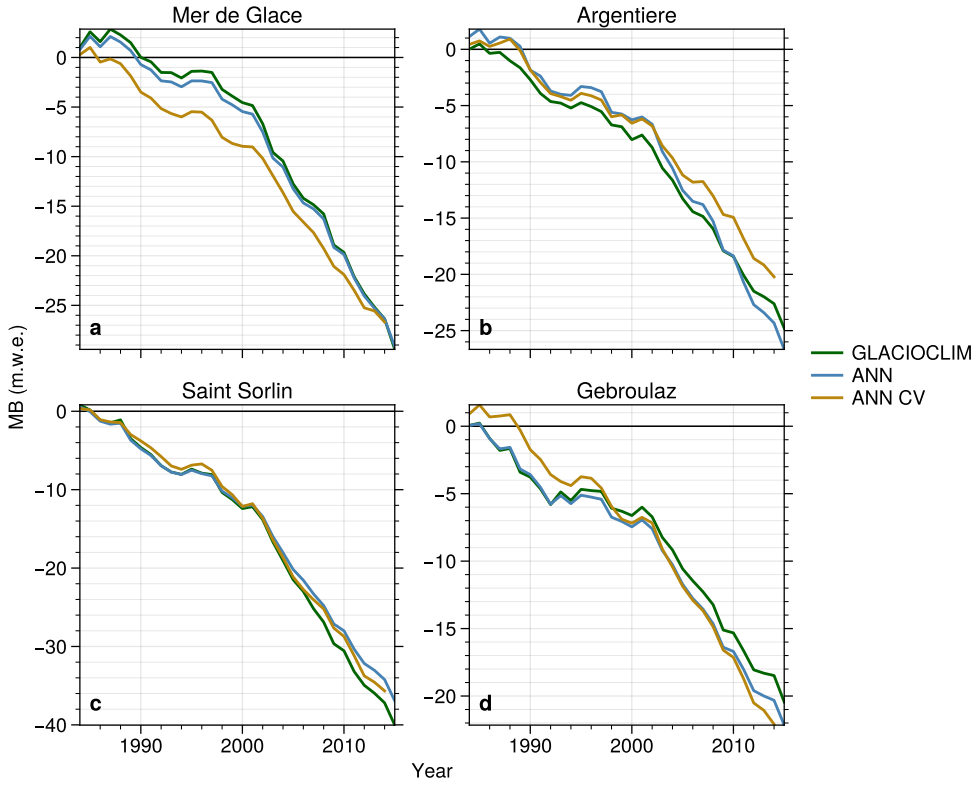

Figure 3.7: Comparison between glaciological observations from the GLACIOCLIM observatory, cross-validated MB reconstructions from this study (ANN CV) and fitted MB reconstructions (ANN). The cross-validated models are shown to display the out-of-sample performance. The fitted reconstructions display the actual reconstructions from the dataset, with models especially fitted for glaciers with data.

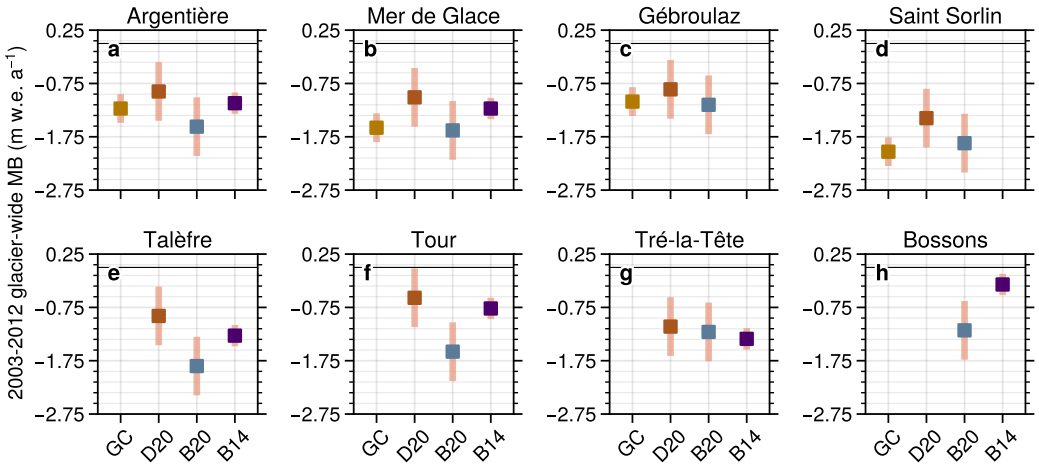

Figure 3.8: Comparison between glaciological observations from the GLACIOCLIM observatory (GC), ASTER geodetic mass balances from Davaze et al. (2020) (D20), the deep learning reconstructions from the present study (B20) and Pléiades geodetic mass balances from Berthier et al. (2014) (B14).

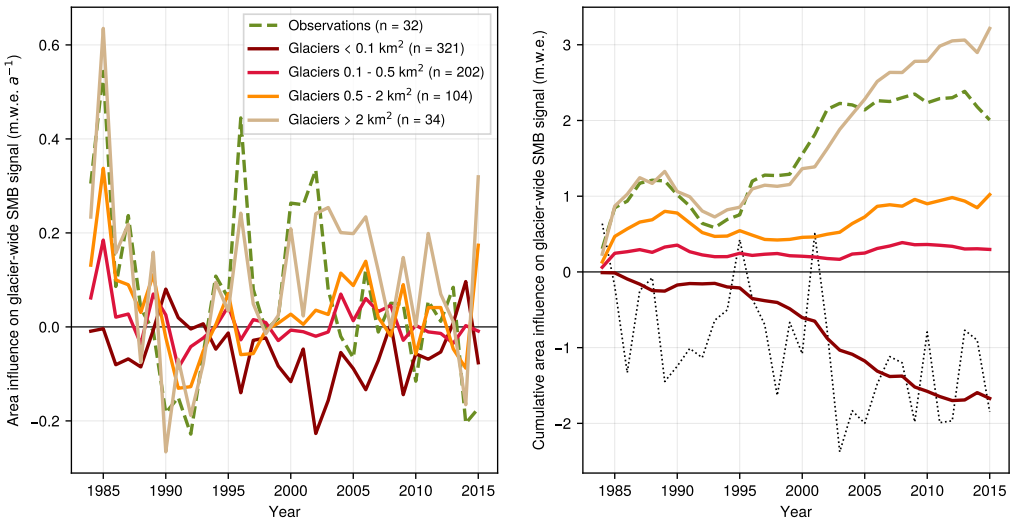

Figure 3.9: Influence of glacier area on the glacier-wide MB signal. The reconstructed median annual glacier-wide MB of the 661 glaciers in the French Alps can be seen as a proxy of the climate signal in the region. It is subtracted to the mean annual glacier-wide MB of the glaciers with observations and to four different subsets of reconstructions divided into glacier area size, showing only the annual differences based on glacier area classes. The dotted line depicts the subtracted signal (non cumulative) in order to give some context.

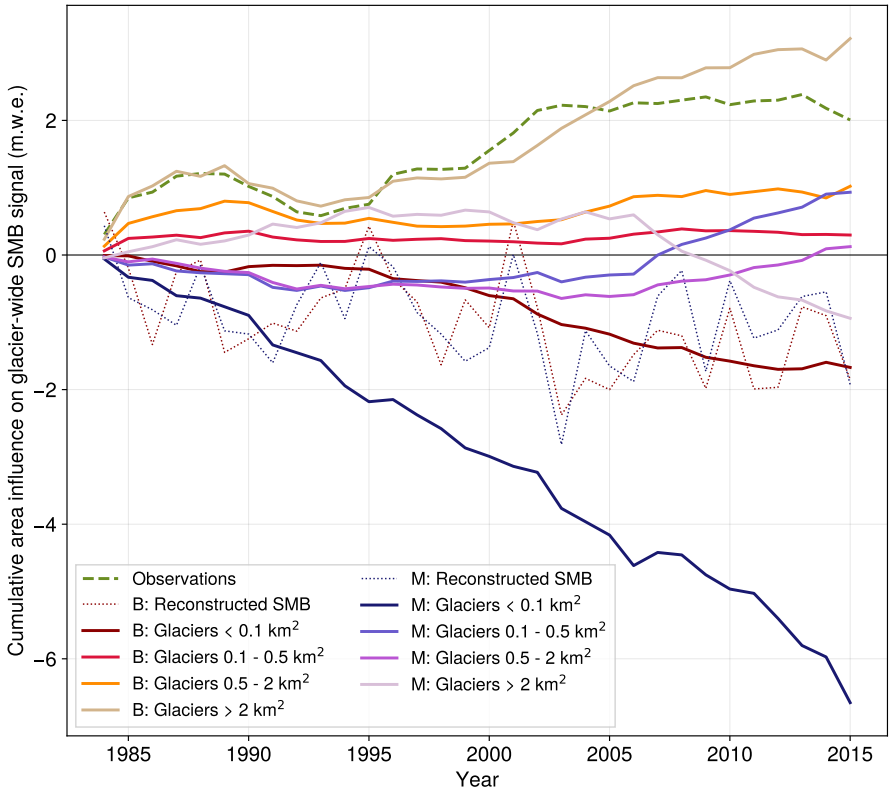

Figure 3.10: Same as S3 but comparing this study to the updated version of Marzeion et al. (2015). In the legend, "B" stands for Bolibar et al. (this study) and "M" for the update of Marzeion et al. (2015). Both models show a relatively similar gradient effect with respect to glacier area, with differences in the amplitude of the effects. The main differences appear from 2007, where small and middle sized glaciers (0.1 - 2 km^2) from the update of Marzeion et al. (2015) switch to a positive influence, as opposite to large glaciers (> 2 km^2), which transition to a negative influence. The reconstructed MB dotted lines are not cumulative and they are depicted in order to give some context of the subtracted climate signal.

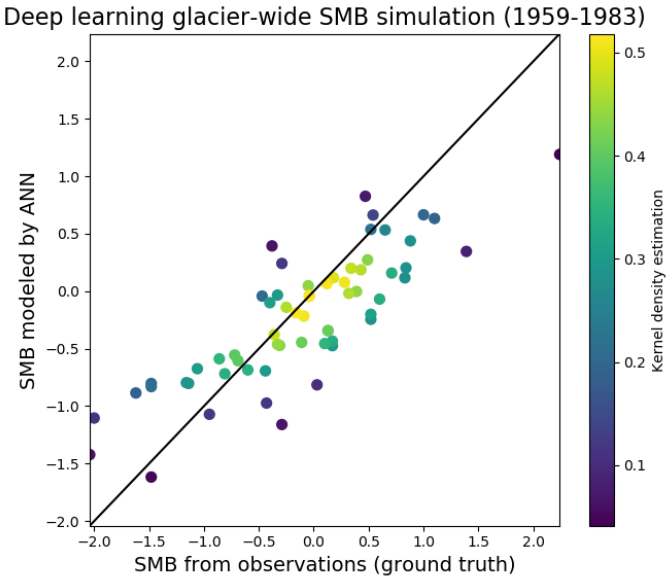

Figure 3.11: Cross-validation for annual glacier-wide MB values outside the main 1984–2014 training period. The black line indicates the one-to-one reference. Simulations have been done from 1959, the earliest date with observations to validate against the maximum number of values. This serves to confirm that the model is capable of reproducing glacier-wide MB outside the main observed period.

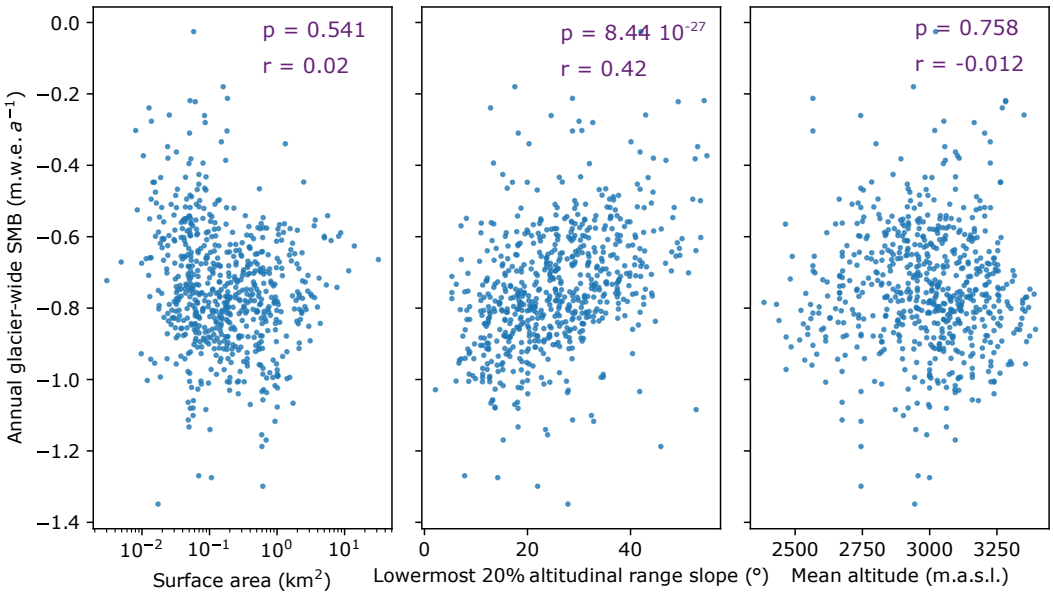

Figure 3.12: Average annual glacier-wide MB for each glacier over the entire study period with respect to (a) glacier surface area, (b) the lowermost 20% altitudinal range slope and (c) mean glacier altitude. p indicates the p-value and r the correlation between the topographical variables and the average glacier-wide MB.

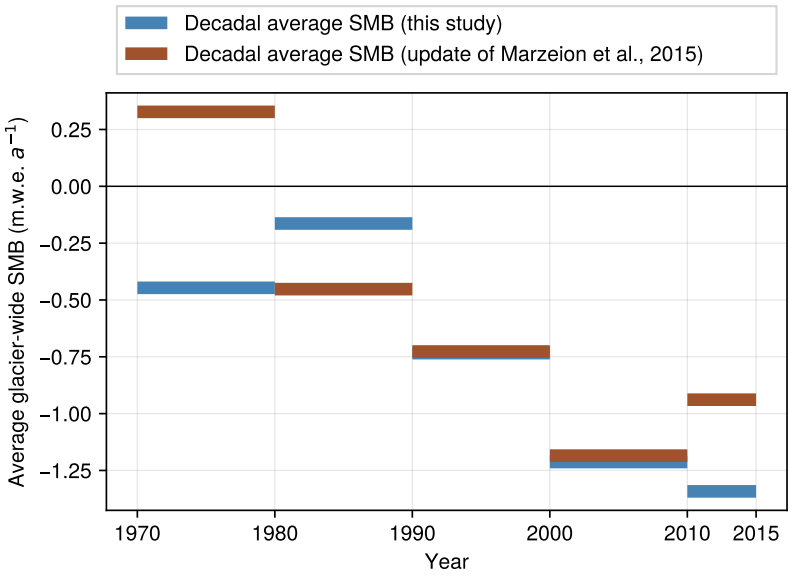

Figure 3.13: Comparison of area-weighted decadal glacier-wide MB simulations in the French Alps between this study and an update from Marzeion et al. (2015).

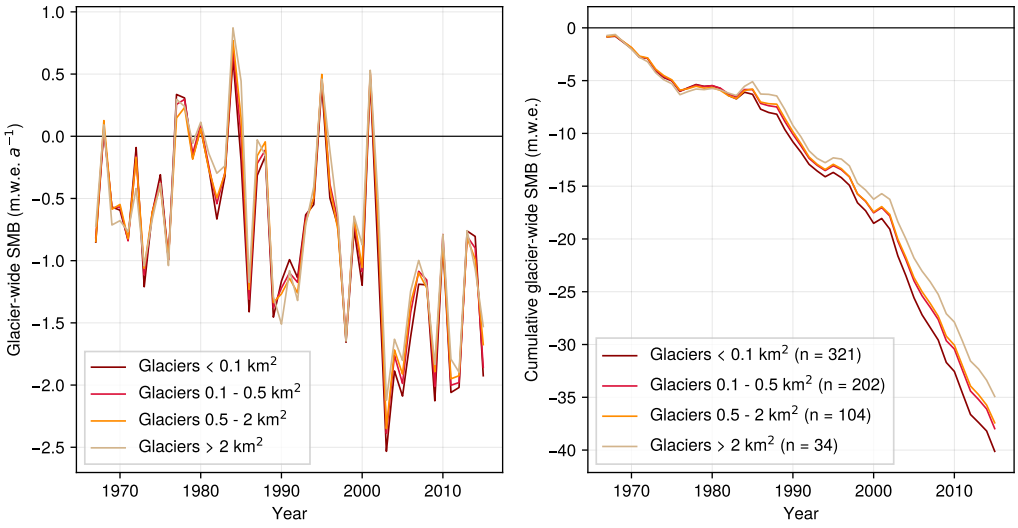

Figure 3.14: Average annual glacier-wide MB per glacier area classes

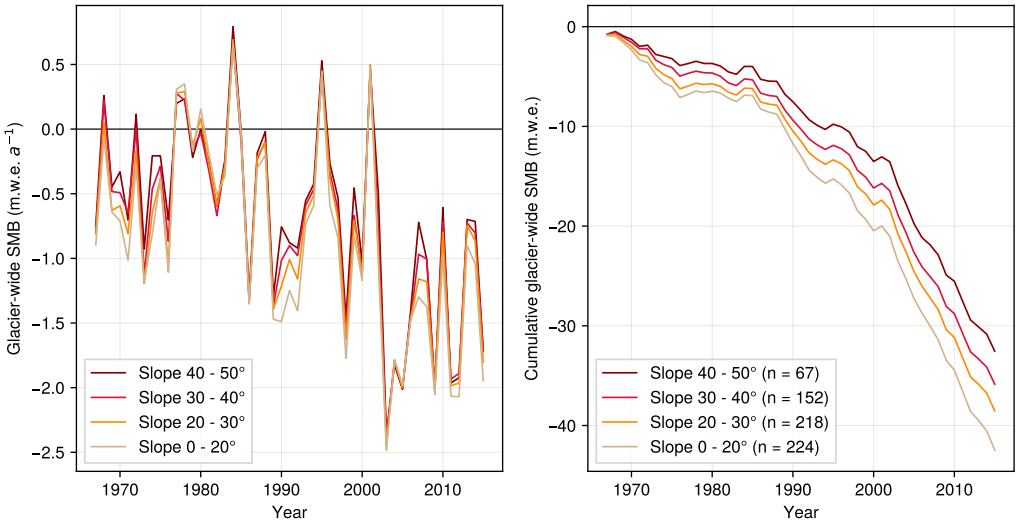

Figure 3.15: Average annual glacier-wide MB for classes of glacier slope of the lowermost 20% altitudinal range (i.e. a proxy of the glacier's tongue slope)

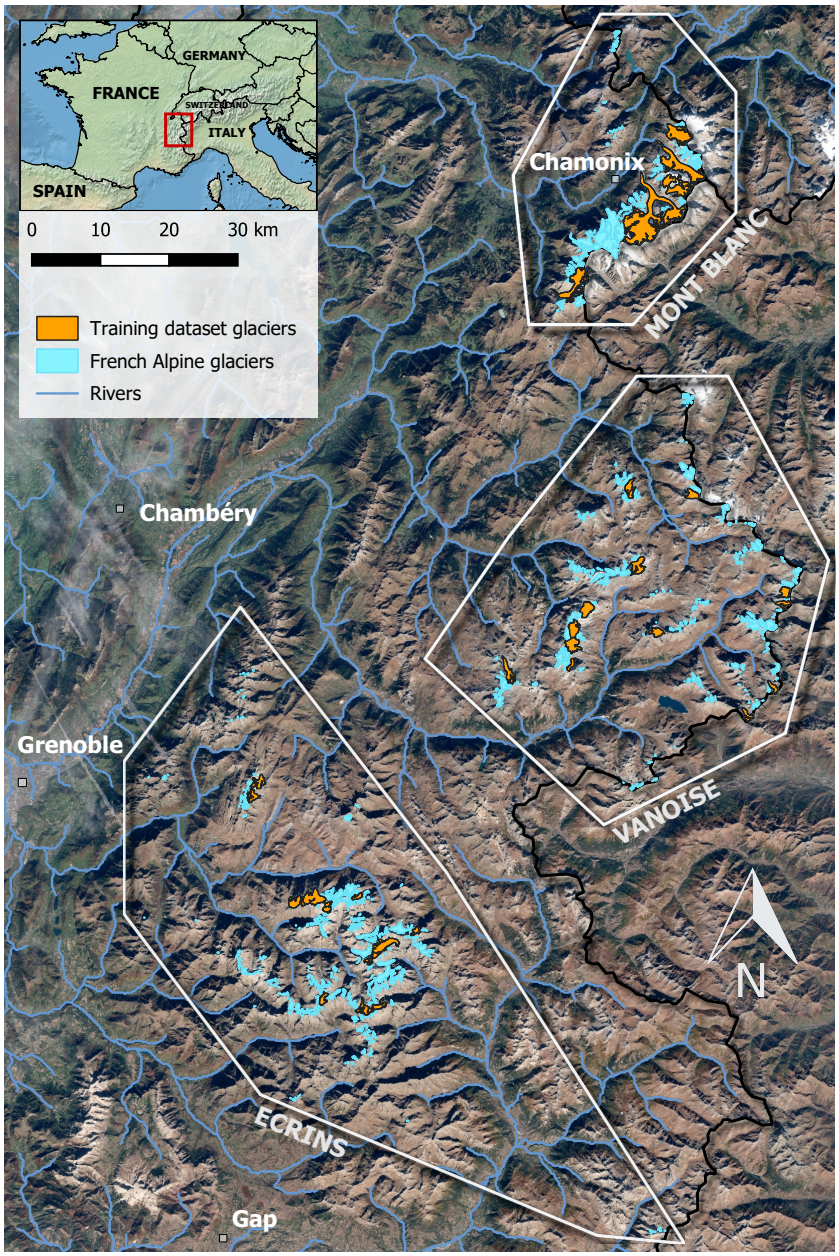

Figure 3.16: French Alpine glaciers used for model training and validation and their classification into three clusters or regions (Écrins, Vanoise, Mont-Blanc). Coordinates of bottom left map corner: 44°32' N, 5°40' E. Coordinates of the top right map corner: 46°08' N, 7°17' E.