Using an ensemble of artificial neural networks to convert snow depth to snow water equivalent over Canada

Konstantin Franz Fotios Ntokas[1,2], Jean Odry[1], Marie-Amélie Boucher[1], and Camille Garnaud[3]

[1]Université de Sherbrooke, Department of Civil and Building Engineering, Sherbrooke, Canada
[2]Technische Universität Berlin, Berlin, Germany
[3]Environment and Climate Change Canada, Dorval, Canada

Correspondence: Konstantin Ntokas (konstantin.ntokas@yahoo.de)

Abstract. Canada's water cycle is driven mainly by snowmelt. Snow water equivalent (SWE) is the snow-related variable that is most commonly used in hydrology, as it expresses the total quantity of water (solid and liquid) stored in the snowpack. Measurements of SWE are, however, expensive and not continuously accessible in real time. This motivates a search for alternative ways of estimating SWE from measurements that are more widely available and continuous over time. SWE can be calculated by multiplying snow depth with the bulk density of the snowpack. Regression models proposed in the literature first estimate snow density and then calculate SWE. More recently, a novel approach to this problem has been developed and is based on an ensemble of multilayer perceptrons (MLPs). Although this approach compared favourably with existing regression models, snow density values at the lower and higher ends of the range remained inaccurate. Here, we improve upon this recent method for determining SWE from snow depth. We show the general applicability of the method through the use of a large data set of 234 779 snow depth-density-SWE records from 2878 non-uniformly distributed sites across Canada. These data cover almost four decades of snowfall. First, it is shown that the direct estimation of SWE produces better results than the estimation of snow density followed by the calculation of SWE. Second, optimizing MLP parameters separately for each snow climate class further improves the accuracy of SWE estimates. A comparison with commonly used regression models reveals that the ensemble of MLPs proposed here leads to noticeably more accurate estimates of SWE. This study thus shows that delving deeper into artificial neural network theory helps improve SWE estimation and that using a greater number of MLP parameters could lead to even further improvements.

1 Introduction

Snowmelt plays a major role in the hydrological cycle of many regions of the world. Casson et al. (2018) determined that snow accumulation and melt are the main drivers of the spring freshet in the Canadian subarctic region, and Pomeroy et al. (2011) demonstrated that over 80% of the annual runoff in the Canadian Prairies is derived from snowmelt. An accurate prediction of

the accumulation and melting of snow is therefore of interest for various applications, including the management of reservoirs for hydroelectric power generation, irrigation and water supply, and climate impact studies.

Consequently, many hydrological models include a snow module to estimate the snow water equivalent (SWE). SWE is
25 of great interest in hydrology because it describes the volume of water stored in the snowpack (e.g. Kirnbauer et al. (1994); Dozier (2011); Hock et al. (2006); Barnett et al. (2005); Seibert et al. (2014)). Kinar and Pomeroy (2015) reviewed techniques for measuring snowpack variables. The existing techniques for determining SWE are either time-intensive, due to manual snow surveys, or cost-intensive because of the necessary and expensive equipment. Sturm et al. (2010) estimated SWE measurements to be 20 times more expensive than snow depth measurements. SWE can also be calculated, however, using snow depth and
30 the volumetric mass density of snow.

Snow depth can be measured inexpensively by ultrasonic distance sensors. Furthermore, light detection and ranging instruments (lidar) installed on aircraft can measure snow depth remotely (e.g. Painter et al., 2016; Kim et al., 2017). Lidar penetrates vegetation, which makes it a very good tool for measuring snow depth in forested environments. These observations also have a high spatial resolution, which makes lidar particularly interesting in mountainous areas where snow depth is highly spatially
35 variable. At present, the high cost of using manned aircraft prevents this technique from becoming operational at a large scale, although Bühler et al. (2016) have demonstrated the potential for unmanned aerial surveys using low-cost, remotely piloted drones to measure snow depth at a large scale and at a high spatial resolution. Lettenmaier et al. (2015) discussed remote sensing techniques in hydrology and concluded that measurement of SWE from space "remains elusive". However, Environment and Climate Change Canada (ECCC) and the Canadian Space Agency (CSA) are currently collaborating on the Terrestrial
40 Snow Mass Mission (Garnaud et al., 2019), a satellite mission that aims to measure SWE with a dual Ku-band radar.

A number of regression models have already been proposed to convert snow depth to SWE (e.g. Jonas et al., 2009; Sturm et al., 2010; Painter et al., 2016; Broxton et al., 2019). In general, these models first estimate snow density from snow depth and then calculate SWE. Sturm et al. (2010) argue that the variation of snow density is four times less than that of snow depth. Therefore, a model that measures the more dynamic parameter (snow depth) and estimates the more conservative parameter
45 (snow density) offers promise. The models are trained on field measurements using regression analysis. Additional geophysical classifications are used to obtain more precise results. For example, Sturm et al. (2010) and Bormann et al. (2013) used the snow classification of Sturm et al. (2009) in which individual regression models were trained for each class. Another example of geophysical classification is given by Jonas et al. (2009). They applied an individual regression model to each month and elevation class to separate respectively the data both temporally and spatially.

50 Physics-based approaches have also been proposed for converting snow depth to SWE. For instance, Painter et al. (2016) used the snowmelt model *iSnobal*, which calculates snow density while incorporating snow ageing, mechanical compaction, and the impact of liquid water with adjustments for deposits of new snow. Because of their complexity, however, physics-based models can be computationally expensive.

A few recent studies have suggested the use of artificial neural networks (ANN) to obtain a relationship between snow depth
55 and snow density, which could then be used to obtain estimates of SWE. Broxton et al. (2019) applied an ANN model to a set of snow measurements at a very high spatial resolution. Their study focused on two approximately $100\,\text{km}^2$ areas in Arizona,

USA, for both mid- and late-winter conditions, during which a total of 300 density-depth tuples were obtained manually. These tuples were used to train the neural network. Their model, which consisted of a simple network structure with one hidden layer containing 10 neurons, also incorporated other physiographic factors obtained from lidar measurements. The Levenberg–
60 Marquardt algorithm optimized the model over a run of 50 epochs. The Broxton et al. (2019) approach allows for the use of high-resolution lidar measurements of snow depth; the ANN model then converts these measurements to snow density and subsequently to SWE to produce improved maps of depth, SWE, and snow density at a very high resolution.

Odry et al. (2020) also applied ANN to estimate snow density from snow depth, but they focused on developing a method that would be applicable over a very large spatial extent. In their study, they used almost $40\,000$ measurements from approxi-
65 mately 400 non-uniformly distributed sites across the province of Quebec, Canada. This study covered a period of 45 years. In contrast to Broxton et al. (2019), the available snow measurements were spatially distant and temporally irregular. The ANN incorporated meteorological data as additional explanatory variables to support the estimation of density. Odry et al. (2020) used an ensemble of multilayer perceptrons (MLPs, a type of ANN) to provide estimates, at least in part, of the uncertainty associated with converting snow depth to density. A comparison of the ANN model of Odry et al. (2020) with the regression
70 models of Jonas et al. (2009) and Sturm et al. (2010) in a leave-one-out set-up showed that the ANN model provided the most accurate results. However, Odry et al. (2020) also noted that all three models performed relatively poorly for very low or very high snow densities. Given that high densities generally correspond to the beginning of the melt period, it is particularly crucial to obtain accurate SWE values for this period of the year.

This study is a follow-up to the work of Odry et al. (2020). Here we test means of improving their initial model. Specifically,
75 we verify two hypotheses: (1) using SWE instead of density as the target variable for the ANN will produce more accurate estimates of SWE; (2) in-depth testing of different ANN structures will improve estimates of SWE. We also take the opportunity to apply the proposed snow depth–SWE conversion framework to the entire area of Canada and test its applicability in a broader context. Canada is characterized by a range of climates associated with different snow classes.

The remainder of the paper is organized as follows. In Sect. 2, we present the main background elements and the essential
80 literature review. This review includes the mathematical theory of MLPs, the applied snow climate classes (Sturm et al., 2009), both statistical regression models used for comparison against our neural network models, and the performance assessment metrics. Section 3 presents the experimental protocol, including the available data and input variables and the procedure for determining the MLP architecture. Results are presented in Sect. 4, and our conclusions are summarized in Sect. 5.

2   Literature review

2.1   The multilayer perceptron as a basic tool

Following Goodfellow et al. (2016), a multilayer perceptron (MLP) is a fully connected, feed-forward artificial neural network, meaning that information can only travel in one direction within the network. Being fully connected entails that no initial assumptions need to be made in regard to the data.

The goal is to approximate a desired function f such that $y = f(\boldsymbol{\theta}, \boldsymbol{x})$, where y is the target variable (in our case SWE), $\boldsymbol{x}$ represents the input variables, and $\boldsymbol{\theta}$ represents some parameters, namely weights and biases. The use of non-linear activation functions in each neuron enables the network to approximate non-linear functions f.

To determine the parameters, one can perform an optimization over the space of the weights and biases by using a training data set. To do so, the parameters are initialized commonly at random and close to zero by following either a uniform or Gaussian distribution. For regression problems, the mean square error (MSE) is used as the objective function in the optimization. Goodfellow et al. (2016) show a derivation of the MSE from the maximum likelihood estimator. During training, the training data set is presented multiple times to the model, and the term *epoch* is used to denote one run over the entire training data set. It is recommended that between each epoch the order of the data points of the training set is permuted. This prevents the model from learning features in the order of insertion. The training of a MLP is related to an ordinary optimization. The difference is that we optimize directly for a training data set with the goal that the measure of performance of a validation data set is also optimized. Commonly, there will be a turning point where the training error continues to decrease, but the validation error starts to increase. For a single deterministic MLP, this turning point would be chosen as the best number of epochs for the model. When training an ensemble of MLP, however, it is ideal to maintain diversity among the ensemble members to cover the range of uncertainty pertaining to the target variable. If all members of the ensemble were to be trained until the aforementioned turning point, they would all become very similar, and it would become pointless to use an ensemble in the first place. For this reason, we adopt here the protocol proposed by Boucher et al. (2010) and monitor rather the performance of the entire ensemble at each epoch (see Sect. 4.1).

During the optimization, a non-convex function is solved; this function contains multiple minima having a similar performance. The objective function with respect to the parameters is non-convex because of several symmetric configurations of a neural network. Thus, exchanging the associated bias and weights of one neuron with another neuron in the same layer entails the same results. Furthermore, ANN applications usually use input variables that are related to each other. The interchangeability of dependent input variables results in multiple parameter sets having a similar performance. The number of dimensions in the optimization is equal to the number of parameters, and it is also argued by Goodfellow et al. (2016) that local minima are rare in high-dimensional spaces. Dauphin et al. (2014) point out that with increasing dimensions, the ratio of the number of saddle points to the number of local minima increases exponentially. Therefore, an optimization method is needed that does not become stuck in saddle points.

The concept of stochastic gradient methods has been introduced to avoid saddle points. These methods are based on the ordinary steepest gradient method; however, rather than using the entire training data set (batch) at once, a number of data records (or "minibatches") are taken for each iteration during the optimization. Consequently, the optimization surface is slightly altered at each iteration. This can even sometimes help to escape shallow local minima. For the interested reader, Goodfellow et al. (2016) provide an overview of various optimization algorithms.

The related studies of Broxton et al. (2019) and Odry et al. (2020) both used the second-order Levenberg–Marquardt algorithm. Broxton et al. (2019) worked with the *Matlab deep learning toolbox* and Odry et al. (2020) used an in-house built *Matlab* toolbox by Dr Kris Villez, derived from the *Matlab deep learning toolbox*. The conversion of the codes into *Python* revealed

that the Levenberg–Marquardt algorithm has an oscillating convergence of the training error because the minimum is over-
125   jumped multiple times. This results in a higher computational cost. The *Matlab deep learning toolbox* only shows successful
parameter updates where the error of the objective function is decreased. This conceals multiple failures. Further discussion
about the drawbacks of the *Matlab toolbox* is available in Kwak et al. (2011). Because of the exponentially increasing dimen-
sions of optimization with larger neural networks, we will only consider first-order optimizers, as these are computationally
more efficient. Goodfellow et al. (2016), in their overview of optimization methods, conclude that stochastic gradient methods
130   with adaptive learning rates are the best choice for a first-order optimizer. Schaul et al. (2014) compare stochastic gradient
methods by testing them on small-scale problems. The algorithms RMSProp and AdaDelta produce good results. The draw-
back of RMSProp is the need for a global learning rate, which must be defined by the user. To address this problem, Zeiler
(2012) introduced the algorithm AdaDelta, which eliminates the requirement of this global learning rate.

There are several reasons that support the use of an ensemble rather than a single model. First, random parameter initial-
135   ization can end up in different local minima on the parameter surface with similar performance. This situation is related to
the concept of equifinality, introduced by Bertalanffy (1968) for open systems stemming from the work of the biologist Hans
Driesch. Therefore, using an ensemble accounts for the uncertainty of the model parameters. Second, the ensemble offers the
possibility of probabilistic simulations. Therefore, probabilistic evaluation methods, introduced in Sect. 2.4, can be used. This
leads to greater insight into model performance. Third, the different members of the ensemble can be used in a Kalman filter
140   or a particle filter for data assimilation in a hydrological model.

Finally, in regard to the architecture of the MLP, through the universal approximation theorem, it is proven that any contin-
uous function can be approximated by a feed-forward neural network with a single hidden layer under mild assumptions. This
theorem was first proven by Cybenko (1989) for the sigmoid activation function. Hornik (1991) showed the same theorem is
independent of the choice of the activation function but assumes that the output layer is linear, which is the case for regression
145   problems. The theorem does not make any claims about guidelines of the architecture or about the learning ability of the model.
In ANN applications, therefore, the perfect approximation of the function is not obtained. Rather, there is a trade-off between
computational cost, the time to test different architecture compositions, and accurate approximation.

According to Goodfellow et al. (2016), the most common activation functions are the hyperbolic tangent (tanh) and the
rectified linear unit (ReLU) functions. The calculation of the derivative of tanh is computationally expensive, and the derivative
150   of the tanh function for large values is almost zero, which slows down learning. The derivative of the ReLU function is easy
to compute, and the derivative stays constant for large positive values. However, the derivative is zero for negative values,
which stops the training within the optimization. This phenomenon is known as "dying neurons". The Leaky ReLU function
can tackle this issue by assigning a slightly positive slope, typically 0.01, to the negative part of the ReLU function.

2.2 Snow classification

155   Because of the very large territory covered by this study, and in accordance with the findings of Sturm et al. (2010) and
Bormann et al. (2013), we expect that training a different ensemble of MLP for each snow class will result in an improved
accuracy for estimates of SWE compared to training an ensemble over the data for Canada as a whole.
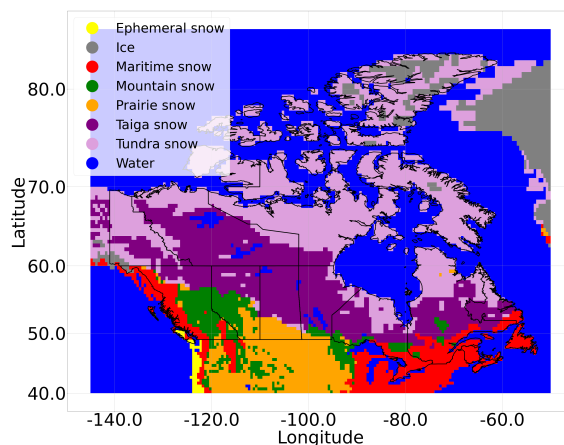
Figure 1. Snow classes across Canada as defined by Sturm et al. (2009)

Sturm et al. (2009) propose a global seasonal snow classification system that separates the world map into seven different snow classifications at a resolution of $0.5° \times 0.5°$. Snow can vary in character (e.g. snow found in areas close to coasts consists

160   mainly of wet snow, whereas snow located in the continental interior is characterized by dry snow). Therefore, the data within one snow class should be systematically more consistent and may show less variability than taking the snow data as a whole. This can help in our study, as the MLP must capture fewer features, which may lead to a better optimization and, in turn, improved performance. The snow classes across Canada are presented in Fig. 1.

Because of the coarse resolution of the snow classification map, some stations close to the coast are classified as *water*.

165   Furthermore, the snow class *ice* only contains 124 records. For these two classes, we assign them to the nearest—in terms of physical distance—accepted snow class.

The empirical cumulative distribution function (ECDF) of snow depth, SWE, and snow density for each snow class is depicted in Fig. 2. The number of records in each snow class is presented in Fig. 2b. The *ephemeral snow* class is found only on Vancouver Island, which also explains the few records of this class. Furthermore, the high annual precipitation and marked

170   variability in elevation across Vancouver Island explains the very high variability of snow depth, SWE, and snow density for this class. When comparing the *taiga* and *tundra snow* classes, one can observe that snow depth is lower in the *tundra snow* class owing to the lower total precipitation in northern Canada. Nonetheless, the density distribution of the *tundra snow* class is higher; this may reflect the wind densification effect in this region as it is relatively more important than other processes for the *tundra snow* class. The *maritime*, *mountain*, and *ephemeral snow* classes have a higher density distribution than the *prairie*

175   and *taiga snow* classes because of the greater oceanic influence on the *maritime* and *ephemeral* regions. This influence results in wetter snow and, therefore, higher snow densities within these latter regions. Most records of the *mountain snow* class are found in the Rocky Mountains along the border of British Columbia and Alberta, an area influenced by the Pacific Ocean, thereby producing a higher density snow. A more thorough analysis of the data revealed that the high number of records in the *maritime* and *mountain* snow classes is due to records from snow pillows in British Columbia, collected between 1996 and
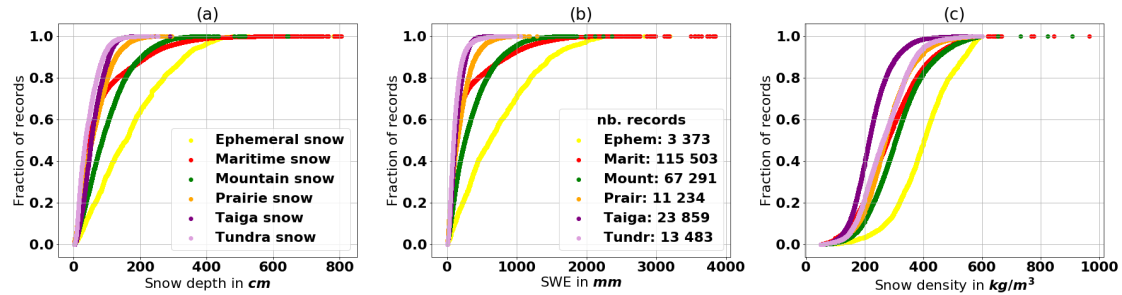
Figure 2. The empirical cumulative distribution function (ECDF) of the variables snow depth, SWE, and density for each snow class

2011. Snow pillows continuously measure the weight of the snow, which is then presented as a daily measurement for some stations in this data set. British Columbia contains almost half of all snow data records for Canada.

2.3 Regression models

In this section we introduce two regression models from the literature, which we use as benchmarks to compare with our MLP-based conversion model, using the performance indicators described in Sect. 2.4.

2.3.1 The Sturm model

Sturm et al. (2010) proposed a model to convert snow depth into bulk density for the snow cover. The estimated density is then used to calculate SWE. As observed by Dickinson and Whiteley (1972), the range of snow density is smaller than the range of snow depth. Therefore, Sturm et al. (2010) claim that estimating the snow density is probably the most accurate way to estimate SWE. The regression model is formulated as:

$$\rho_{sim} = (\rho_{max} - \rho_0)[1 - e^{-k_1 SD_{obs} - k_2 DOY_{obs}}] + \rho_0, \tag{1}$$

where $[\rho_{max}, \rho_0, k_1, k_2]$ are parameters, SD_{obs} is the snow depth, and DOY_{obs} is the number of days since 01 October. The value of DOY_{obs} is 0 on 01 January. A different regression model is run for each snow class. The snow classes are defined by Sturm et al. (2009). The parameters for each snow class are obtained by carrying out an optimization on a training data set associated with that snow class. The RMSE between the estimated and the measured snow density is used as an objective function in the optimization.

2.3.2 The Jonas model

Jonas et al. (2009) propose a simple linear regression model that uses elevation classes to separate the data spatially and divide the data into months. This model also estimates density, and from this density, it calculates the SWE. The regression is defined

as:

$$\rho_{sim} = a\, SD_{obs} + b + offset_{reg},\qquad(2)$$

where a and b are the parameters of the linear regression, SD_{obs} is the observed snow depth, and $offset_{reg}$ is a regional specific parameter. The data set is split by month and into three elevation classes, denoted by $x < 1400\,\mathrm{m}$, $1400\,\mathrm{m} \leq x < 2000\,\mathrm{m}$, and $x \geq 2000\,\mathrm{m}$, where x is the elevation of the site. After that, the parameters a and b are derived by fitting a linear regression to each portion of the data set. This process produces 36 independent linear regression models. After solving for the parameters of the linear regressions, we perform a simulation without $offset_{reg}$. The regional specific parameter $offset_{reg}$ is thus the average of the model residual between the simulated and observed density of samples in a given region. In our study, the snow classes defined by Sturm et al. (2009) are used as these regions. Note that the separation into the snow classes for the $offset_{reg}$ calculation is applied independently after the linear regressions are performed for each month and elevation class. The $offset_{reg}$ parameter eliminates the regional bias.

2.4 Model evaluation

Gneiting and Raftery (2007) evaluate the quality of ensemble simulation by scoring rules. A scoring rule quantifies the quality on the basis of the predicted distribution of the ensemble and the observation. A strictly proper scoring rule is defined such that it has a unique global minimum, which is the distribution of the observation itself. This minimum shows the highest possible performance of the simulation. The probability density function (pdf) is usually derived by distribution fitting to the members of the ensemble or by simply taking the empirical cumulative distribution function (cdf) over the members and deriving the pdf from this. In this study, only the empirical case is used to eliminate the uncertainty added by the distribution fitting.

Furthermore, an ensemble is said to be reliable if the relative frequency of the event, for a given simulation probability, is equal to the simulation probability.

2.4.1 The ignorance score

Roulston and Smith (2002) introduced the ignorance score, which is defined as:

$$IGS(f_M, o) = -log_2(f_M(o)),\qquad(3)$$

where f_M is the simulated pdf, and $o \in \mathbb{R}$ is the observation. Since $f_M(o) \in [0,1]\, \forall o \in \mathbb{R}$, the ignorance score takes values within $\mathbb{R}_0^+$. Furthermore, it is not defined for $f_M(o) = 0$. It assigns the minimum value of zero if the probabilistic forecast predicts the observation with a probability equal to 1, indicating a perfect simulation. Therefore, the ignorance score is a strictly proper scoring rule. To assess the model, the average of the ignorance scores over all n records is taken, defined as:

$$\overline{IGS} = \frac{1}{n}\sum_{j=1}^{n} -log_2(f_{Mj}(o_j)),\qquad(4)$$

where f_{Mj} is the pdf derived from the ensemble of the jth simulation, and o_j is the jth observation.

In the empirical case, i.e. without distribution fitting, we consider the sorted predicted ensemble with m members by $\{s_1, s_2, ..., s_m\}$. The cumulative distribution function is simply the staircase function over the members. From this, we construct

230 a pdf by assigning each area between two members:

$$f_M(x) = \frac{1}{m(s_i - s_{i-1})} \quad \forall x \in [s_{i-1}, s_i].$$

If the observation coincides with a member, the larger probability of the two adjacent areas is taken. Furthermore, $f_M(x)$ is set to 0.001, if the observation lies outside of the ensemble.

2.4.2 The continuous ranked probability score

235 Given that the ignorance score evaluates the simulated probability function only at the point of observation, no information about the area surrounding the observation or the shape of the probability function is included. The continuous ranked probability score (CRPS) addresses this drawback by working directly on the cdf. The CRPS of one record is defined as:

$$CRPS(F_M, o) = \int_{-\infty}^{\infty} (F_M(w) - \mathbf{1}_{[o,\infty)}(w))^2 dw, \tag{5}$$

where F_M is the cdf derived from the ensemble, and $o \in \mathbb{R}$ is the observation. The integrand is the squared difference between
240 the cdf derived by the ensemble and the cdf derived by the observation. The CRPS takes values in $\mathbb{R}_0^+$, where zero is assigned if the predicted cdf is equal to the cdf derived by the observation. This is the perfect simulation. This describes the CRPS as a strictly proper scoring rule. When calculating the empirical CRPS, the staircase function over the members within the ensemble is taken as the cdf.

Hersbach (2000) presents a decomposition of the empirical CRPS into a reliability part and the potential CRPS. We denote
245 the ensemble of one record by $\{x_1, ..., x_m\}$. The reliability part is then defined as:

$$Reli = \sum_{i=0}^{m} g_i \left(o_i - p_i\right)^2, \tag{6}$$

where o_i is the relative frequency of the observation being smaller than the midpoint of the bin $[x_i, x_{i+1}]$, p_i is $\frac{i}{m}$, the value of the cdf of the ensemble at x_i, and g_i is the average bin size of $[x_i, x_{i+1}]$. To get the reliability score of the model, the average over all records is taken.

250 Small values of the reliability portion testify to the high reliability of the ensemble simulation. The reliability portion is highly related to the rank histogram, introduced in the next section (Sect. 2.4.3). The difference is that the reliability part takes the spread of the ensemble into account, which is not incorporated into the rank histogram. When the reliability part is subtracted from the empirical CRPS, the potential CRPS remains. The potential CRPS is the same as the CRPS when the reliability part is equal to zero, which indicates a perfectly reliable model. The potential CRPS is related to the average spread
255 of the ensemble. A narrow ensemble leads to a small potential CRPS.

2.4.3 The rank histogram

The rank histogram was developed independently and almost simultaneously by Anderson (1996), Hamill and Colucci (1997), and Talagrand et al. (1997). It determines the rank of each observation within the associated predicted ensemble. The ranks of all observations are then presented in a histogram. A perfectly reliable ensemble forecast shows a flat rank histogram. Furthermore,

260 the rank histogram can reveal information about bias and under- and overfitting. For instance, if the rank histogram shows a decreasing trend with higher ranks, the model has a positive bias. An increasing trend with higher ranks shows a negative bias. Furthermore, a u-shaped rank histogram indicates a too-small ensemble spread. This limited spread is related to an overconfident model, which means that the model ignores some parts of the uncertainty. Several reasons explain this, including the fit of the model. If a model is trained too long, it is trained too specifically on the training data. Consequently, the model

265 performs well on the training data, but it lacks generality, meaning that the model is incapable of simulating an unseen data set. In contrast, a bell shape indicates a too-large ensemble spread. Note that a flat histogram is a necessary condition for a model to be reliable, but this is not a sufficient condition. Examples and further discussion are given by Hamill (2001).

2.4.4 The reliability diagram

The reliability diagram is a visual tool for determining the reliability of an ensemble forecast. The relative frequency of the

270 observation, given the probability of the ensemble, is plotted against the probability of the ensemble.

To construct a reliability diagram, we work with the empirical staircase function over the ensemble members as cdf. Let us assume that we have k bins, denoted by $b_1, ..., b_k$. These bins represent a probability and, therefore, their sizes are $bs_1, ..., bs_k \in (0, 1]$. Furthermore, all bins are centred on 0.5, therefore indicating an interval around the median. Thus, a bin b_i is indicated by the interval $[s_i, e_i]$. We find the s_i- and e_i-quantiles within the ensemble distribution, denoted by s_{en_i} and e_{en_i},

275 respectively. Interpolation is used if s_i and e_i lie between the steps. Therefore, the interval $[s_{en_i}, e_{en_i}]$ shows the interval around the median of the ensemble distribution, which has the probability equal to bs_i. Then, we determine the relative frequency of the observations found within this interval, denoted by o_i. The points (bs_i, o_i) are plotted against the line of the identity function. If all points fall on that line, the model is perfectly reliable, meaning that the relative frequency of the observation is equal to the simulated frequency.

280 ### 2.4.5 Skill scores

Skill scores enable comparing simulated model outputs of different magnitudes. The skill score SS is defined as:

$$SS = \frac{score_s - score_{ref}}{score_{perf} - score_{ref}}, \tag{7}$$

where $score_s$ is the score of the actual simulation, $score_{ref}$ is the score of a reference simulation, and $score_{perf}$ is the score of a perfect simulation. Climatology or persistence is often used as a reference simulation. A skill score takes values from

285 $(-\infty, 1]$. A value of 1 indicates that the actual simulation is a perfect simulation. A negative skill score indicates that the actual

simulation is less accurate than the reference simulation.

Because of the lack of continuous SWE measurements, a variation of climatology is taken as a reference simulation. Records from the training and validation data sets at the same location and within a time window of ±15 days around the date of the 290 corresponding observation are used to build up a reference ensemble having 20 members. For 77 % of the testing records, an ensemble from climatology could be obtained in this manner. In the calculation of the skill score in Eq. (7), the mean over the individual scores is taken within the fraction. This enables us to use only the portion of records where we found the 20 members to calculate the reference scores.

2.4.6 Sensitivity score

295 Following Olden and Jackson (2002), we set input-output records as references where the output is the 20th, 40th, 60th, and 80th percentiles of the target variable SWE in the validation data set. Because single records can deviate from normality, we take the average of 1000 records around the percentile to maintain the generality of the reference record. Subsequently, one input variable is perturbed at a time by taking uniformly distributed values between the maximum and minimum values in the validation set of the considered input variable. The perturbed inputs are inserted into the network, and the perturbed output 300 is generated. The sensitivity score is then given by the RMSE between the perturbed and reference output, averaged over the four reference percentiles and all members. Therefore, it measures the change in the output of the trained ensemble network associated with the change in one input variable.

3 Experimental protocol

3.1 Data availability

305 Environment and Climate Change Canada (ECCC) through Brown et al. (2019) and Ministère de l'Environnement et de la Lutte contre les Changements Climatiques (MELCC) provided a Canada-wide snow data set, which includes SWE, snow depth, and snow density. For the remainder of this paper, this data set will be denoted as the Canadian historical snow survey (CHSS).

For this study, we use the above-mentioned CHSS snow data, collected from 01 January 1980 to 16 March 2017. It consists of 234 779 measurements from 2878 stations. Figure 3 presents some characteristics of the CHSS data set. The distributions 310 of SWE and SD present a right-skewed gamma distribution. Snow density is almost normally distributed in a range from $50\,\mathrm{kg\,m^{-3}}$ to $600\,\mathrm{kg\,m^{-3}}$ with a few outliers at higher values. This distribution is related to the retrieval of two different data sets. The data from ECCC is bounded by the interval $[50, 600]$ for snow density, reflecting snow having a density of up to $600\,\mathrm{kg\,m^{-3}}$. Above this threshold, snow begins to transform into ice. The higher densities encountered in the MELCC data are due to the presence of an ice layer in the snow pack or to measurement artifacts. Figure 4 shows the location of sites within the 315 CHSS data set.
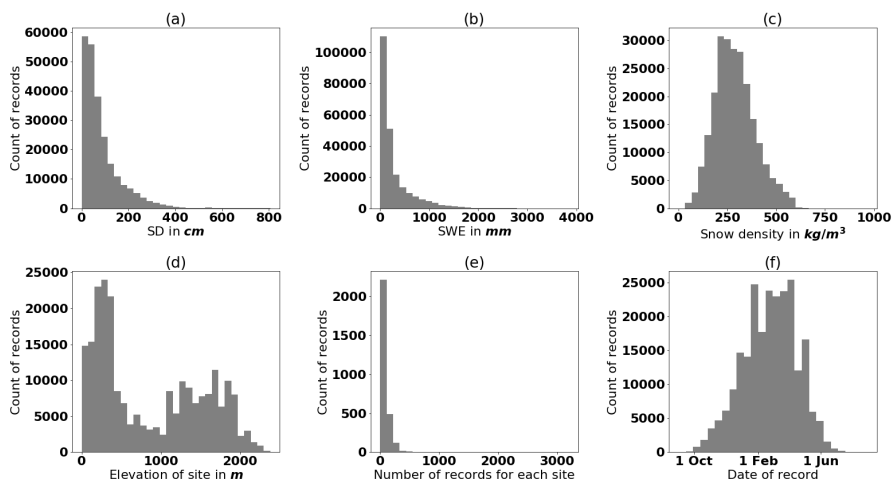
Figure 3. Characteristics of the Canadian historical snow survey (CHSS) data set. The year starts on 01 September to cover the Northern Hemisphere winter season over two calendar years
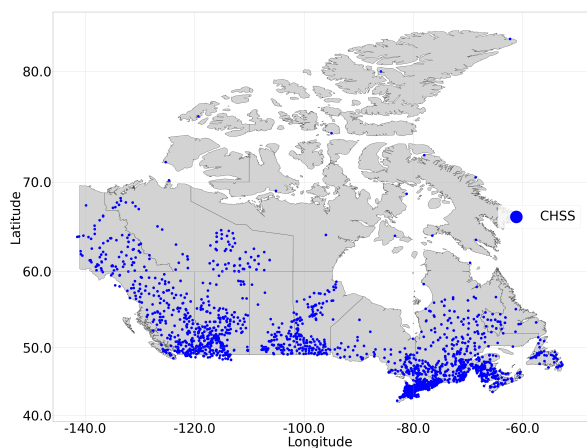
Figure 4. Location of the sites of the Canadian historical snow survey (CHSS) data set

We retrieved daily total precipitation, snow density, as well as the maximum and minimum temperatures for each station of the CHSS from the ERA5 atmospheric reanalyses of the European Centre for Medium-Range Weather Forecasts (ECMWF) provided by the Copernicus Climate Change Service (C3S). Because of the low resolution of the reanalysis grid, we apply a temperature correction, dependent on elevation, to the minimum and maximum temperatures. The modified temperature t_{mod}

320 is defined as:

$$t_{mod} = t_{orig} + lapserate\,(elev_{station} - elev_{grid}), \qquad\qquad (8)$$

where t_{orig} is the original temperature of ERA5, $elev_{station}$ and $elev_{grid}$ are the elevation of the station and the grid point of ERA5 reanalysis, respectively. We apply a constant lapse rate of $-6\,°\mathrm{C}\,\mathrm{km}^{-1}$, a value consistent with studies in the Rocky Mountains (e.g. Dodson and Marks, 1997; Bernier et al., 2011) and the global mean environmental lapse rate of $-6.5\,°\mathrm{C}\,\mathrm{km}^{-1}$ (Barry and Chorley, 1987).

3.2 Explanatory variables

From the total precipitation, temperature, and snow density data, we calculate the following explanatory variables. This initial pool of variables is based on Odry et al. (2020).

- averaged daily snow density (ERA5),
- snow depth,
- number of days since the beginning of winter,
- number of days without snow since the beginning of winter,
- number of freeze-thaw cycles; threshold for freezing and thawing is set at $-1\,°\mathrm{C}$ of the maximum and at $1\,°\mathrm{C}$ of the minimum temperature, respectively
- the degree-day index, i.e. accumulation of positive daily temperatures since the beginning of the winter,
- the snow pack aging index, i.e. the mean number of days since the last snowfall weighted by the total solid precipitation on the day of the snowfall,
- the number of layers in the snowpack estimated from the timeline and intensity of solid precipitation. A new layer is considered to be created if there is a three-day gap since the last snowfall,
- accumulated solid precipitation since the beginning of the winter,
- accumulated solid precipitation during the last n days,
- accumulated total precipitation during the last n days,
- mean average temperature during the last n days (average temperature is taken as the mean of the maximum and minimum temperatures).

We set the beginning of winter as 01 September because the seasonal distribution of the CHSS data set has the first snow records starting from mid-September, with the exception of some outliers. The separation of precipitation into solid and liquid parts is done by:

$$p(t_{av}) = \frac{1}{(1 + e^{-1.54 + 1.24\,t_{av}})}, \tag{9}$$

where $p(t_{av})$ is the probability of snow, dependent on the average temperature t_{av}. Jennings et al. (2018) showed by using precipitation data from the Northern Hemisphere that this logistic regression model outperforms any temperature threshold

Table 1. Variables having the largest absolute Spearman correlation between the target variable and the last three explanatory variables for n ranging between 1 to 10 days

Variable	SWE	Snow density
Accum. solid precipitation in the last	10 days	3 days
Accum. total precipitation in the last	10 days	10 days
Mean average temperature in the last	6 days	7 days

separation model. To determine how many days are considered in the last three explanatory variables, we calculate the Spearman correlation between the target variable SWE or snow density and the range of 1 to 10 days. The results in Table 1, with the other above-mentioned variables, are used as input variables for the network.

3.3 Input uncertainty

355 We target the uncertainty of the measurements of the input variable snow depth. According to the WMO (2014), the error of snow depth measurements should not exceed $\pm 1\,\mathrm{cm}$ if the snow depth is less than $20\,\mathrm{cm}$ and $\pm 5\,\%$ if the snow depth is greater or equal to $20\,\mathrm{cm}$. To give an equal weight to the uncertainty of inputs and parameters, we perturb each record of the training data set 20 times with the above-mentioned variability. The model is then trained on the perturbed data. For the validation, we insert 20 perturbed snow depth values into each trained MLP, resulting in 20 members for each MLP. Therefore, the total

360 ensemble of MLPs produces 400 members. Twenty equally distributed quantiles define the new 20 members. Note that the perturbed training data set is 20 times larger than the number of members. Thus, in one epoch, the optimization algorithm runs over a number of minibatches, which is 20 times the previous number, entailing a smaller numbers of epochs.

3.4 Tested characteristics

After setting up a reference MLP with the characteristics shown in Table 2, one single characteristic is tested at a time. Hereby,

365 all 12 possible explanatory variables from Sect. 3.2 are taken. The number of neurons in the hidden layer is derived by the proposed rules of thumb of Heaton (2008) and Hecht-Nielsen (1989). The number of members is set at 20 because this number showed consistent results during several trials. Table 3 summarizes all the tests that were performed as part of this study to determine the best possible model for converting snow depth to SWE.

Furthermore, the data set is divided randomly into three parts: training, validation, and testing sets (Hastie et al., 2009), each

370 having a proportion of one-third. The training set is used for training the MLPs. The validation set is used to optimize the architecture of the network. The testing set is used for the final model evaluation on an unseen data set and to compare with the regression models in Sect. 4.2 and 4.3, respectively.

Table 2. MLP ensemble characteristics of the reference set-up

Characteristic	Choice
Explanatory variables	12
Number of hidden neurons	10
Target variable	SWE
Activation function in the hidden layer	tanh
Optimization algorithm	AdaDelta
Parameter initialization	$U(-1,1)$
Batch size	100
Shuffling data before each epoch	No
Number of members in the ensemble	20
Input uncertainty of snow depth	No

Table 3. A summary of the tests performed to obtain the final MLP architecture. The parameter initialization [1] is given by Goodfellow et al. (2016), where m indicates the number of inputs, the parameter initialization [2] is given by Glorot and Bengio (2010), where m and n indicate the number of inputs and outputs, respectively. The parameter initialization [3] is our suggestion

Element to test	Options			Results section
Choice of target variable	SWE	Density		4.1.1
Input uncertainty of snow depth	Yes	No		4.1.2
Activation function in the hidden layer	tanh	ReLU	Leaky ReLU	4.1.3
Optimization algorithm	AdaDelta	RMSProp		4.1.3
Shuffling data before each epoch	Yes	No		4.1.3
Parameter initialization	$\sim U\left(-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\right)^{1}$	$\sim U\left(-\sqrt{\frac{6}{m+n}}, \sqrt{\frac{6}{m+n}}\right)^{2}$	$\sim U(-2,2)^{3}$	4.1.3

3.5 Input variable selection

To determine the importance of the input variables, we carry out a sensitivity analysis. In Sect. 2.4.6, the introduced sensitivity
375 score is used to determine the order of the importance of the input variables. Subsequently, a stepwise reduction of the input variables, staring with the least important variable, is performed. The results are shown in Sect. 4.1.4.

3.6 Determination of the number of epochs and the number of hidden neurons

For the last step of the construction of the MLP ensemble, we perform an analysis of number of epochs and number of hidden neurons. An analogous analysis over the number of epochs and hidden neurons is carried out for each snow class to determine
380 a MLP ensemble model for each class.

When adding neurons to the hidden layer, the complexity and the number of parameters (the weights and biases) of the model are increased. This complicates the optimization and therefore, entails a longer training time to obtain the desired parameters.

4 Results

The results are divided into three sections. In Sect. 4.1, we discuss the determination of the MLP architecture, following mainly
385 the outline of Sect. 3. Section 4.2 presents the results of the final MLP model on the testing data set, and in Sect. 4.3 we compare the final MLP model and the regression models introduced in Sect. 2.3.

4.1 Determination of the MLP architecture

4.1.1 Comparison of SWE and snow density as the target variable

A comparison of the two target variables is performed by using the reference MLP architecture presented in Table 2, and Fig.
390 5 presents the results of this comparison. The ensemble of MLPs, with SWE as the target variable, shows slightly better values for the RMSE in Fig. 5d compared with Fig. 5a. Furthermore, in Fig. 5e the reliability part of the CRPS increases later when SWE is the target variable compared with when the target is variable density (Fig. 5b). This behaviour is also consistent with the ignorance score, which has its minimum at epoch 20 in Fig. 5f, whereas in Fig. 5c, the ignorance score for snow density increases from the beginning.
395 This is consistent with the rank histograms in Fig. 6. The higher variability of SWE has a positive effect on the rank histogram, meaning that the spread of the ensemble remains sufficiently large up to epoch ten.

From these results, we can derive that the lower variability of snow density results in too-little spread of the ensemble already after one epoch. This observation allows us to proceed with the target variable SWE. Nonetheless, even with SWE, the spread of the ensemble becomes too narrow, too quickly. This issue is addressed by incorporating input uncertainties, as we show in
400 Sect. 4.1.2.

4.1.2 Input uncertainty

We apply input uncertainty because we cannot train for a sufficient enough period to obtain the best error scores measured by RMSE, MAE, and CRPS. This inability is due to a loss of reliability and overfitting. This type of overfitting is related to the ensemble and does not describe the overfitting of one single MLP. To better understand, we examine the RMSE in Fig.
405 5, which is related to the MSE, the objective function in the optimization of the MLP. The RMSE continues to decrease beyond 15 epochs (optimal trade-off between accuracy and reliability). Therefore, the single MLP remains in an underfit state.
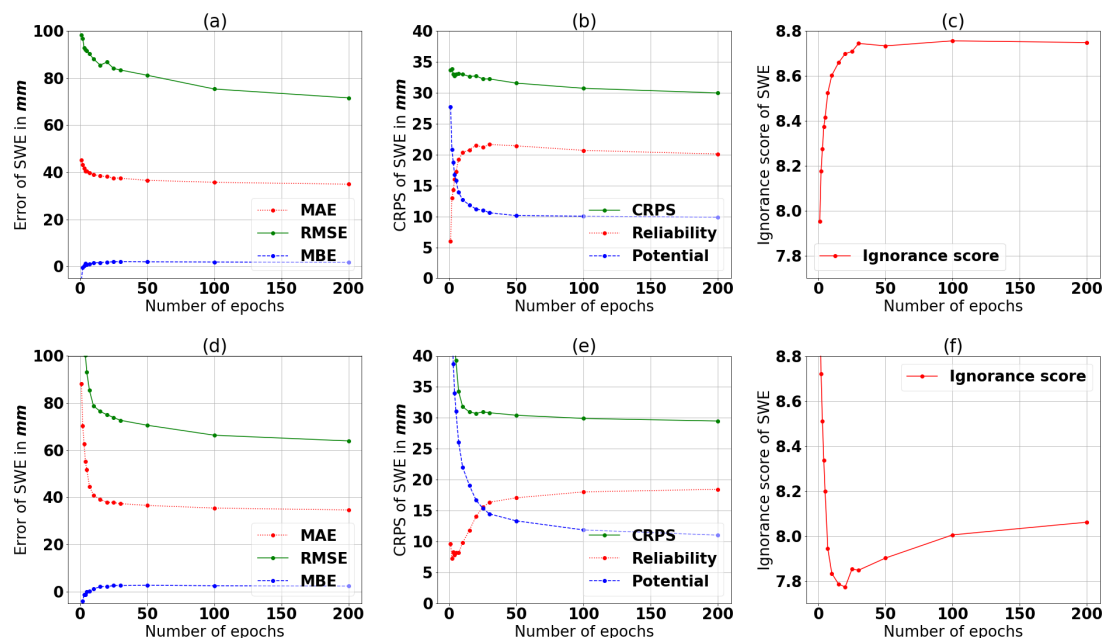
Figure 5. Comparison of (a)–(c) snow density and (d)–(e) SWE as target variables. The scores are evaluated using the validation data set. We use the reference set-up presented in Table 2

Adding regularization would shift the MLP to a more underfit state by simplifying the network. Table 4 shows that incorporating input uncertainty decreases the ignorance score, whereas the other measures are unaffected. Therefore, input uncertainty widens the spread of the ensemble. Further improvements can likely be achieved by incorporating the input uncertainty of the
410 meteorological data through use of the ERA5 ensemble.

4.1.3 Tested characteristics

Table 4 presents all results for the tested characteristics listed in Table 3. The change of the activation function from tanh in the reference set-up to ReLU and Leaky-ReLU does not produce any improvement. Using the ReLu function as the activation function in the network shows a similar but delayed behaviour in the performance scores. In terms of accuracy, we obtain similar
415 results as with the tanh case via a longer learning. Nonetheless, the increase of the reliability part is not delayed as much as the decrease in accuracy. This entails a worse trade-off between accuracy and reliability. When the activation function is set to the Leaky-ReLU activation function, we observe almost no change relative to the ReLU case. The analogous argumentation can be applied.

For the RMSProp algorithm, we set the global learning rate at 0.1, 0.01, 0.001, and 0.0001 during testing. The best result
420 is produced for the trial having a learning rate of 0.0001. The learning rate of 0.001 produces an earlier decrease in accuracy scores and an earlier increase for the reliability part. The trials having a learning rate of 0.1 and 0.01 converge at higher values
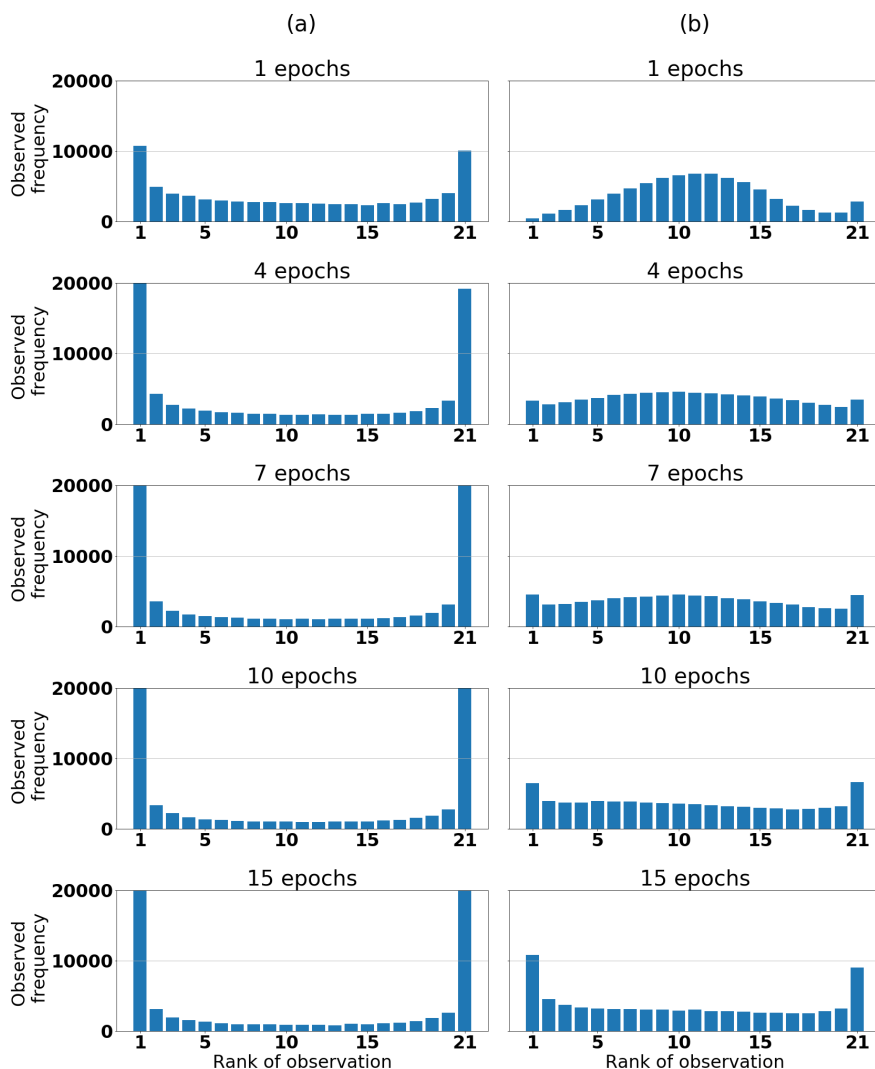
Figure 6. Rank histogram for different numbers of epochs (a) with snow density and (b) SWE as the target variables. We use the reference set-up presented in Table 2

for RMSE and MAE. The higher learning rates can over-jump desirable minima and end up in worse ones. In summary, the overall learning rate must be small enough to achieve the best results. Even lower learning rates will entail a later behaviour. Therefore, the same trade-off between reliability and accuracy can be obtained if training time is increased. When we compare

425 RMSProp to AdaDelta, we note no improvement. Consequently, the AdaDelta method is preferred because there are no learning rates that need to be adjusted.

Table 4. Comparison of all tested characteristics worth considering; Parameter initialisation [2] and [3] as in Table 3; The number of epochs is selected so that the trade-off between reliability and accuracy of the ensemble is increased. The underlined numbers in bold indicate an improvement throughout the testing. ***Combo*** combines para. init.[3], shuffled data, and input uncertainty, and is used as the final set up

	Ref.	Input uncert.	ReLU	Leaky ReLU	RMSProp (rate 0.0001)	Shuffle data	Para. init.[1]	Para. init.[2]	Para. init.[3]	**Combo**
Epochs	15	4	15	10	30	15	25	7	30	**2**
MAE	39.1	39.5	41.4	42.7	40.3	39.0	38.2	43.4	39.7	**37.9**
RMSE	76.5	78.5	82.3	87.3	78.1	75.7	72.1	80.2	79.4	**76.4**
MBE	2.1	1.9	1.7	1.7	1.0	**0.4**	2.8	1.6	1.9	**0.3**
CRPS	30.9	30.9	32.5	33.2	31.6	30.9	33.9	34.4	30.3	**29.3**
Reliability	11.8	11.5	11.7	10.4	10.6	11.9	25.2	13.6	**7.5**	**8.9**
Ign. score	7.79	**7.49**	7.83	7.86	7.81	7.76	8.35	7.94	7.77	**7.4**

A test of the shuffling data produces an almost identical result as that for the reference set-up; however, it eliminates bias, which is almost zero beyond 15 epochs.

In regard to the parameter initialization, the equation [1] in Table 3 is a uniform distribution $U(-0.29, 0.29)$ in our application.
430 The narrow interval results in a too-narrow ensemble, which entails a high value for the reliability part and ignorance score. The second trial, using the parameter initialization [2] (in our case $U(-0.68, 0.68)$), shows a very similar behaviour as the reference set-up but with a lower number of epochs. This matches the interval of the random initialization being more narrow than that in the reference set-up. It entails that the spread of the ensemble narrows more quickly and results in an early increase of the reliability part and an early minimum in the ignorance score at ten epochs. The third trial, using the parameter initialization
435 [3], shows a delayed behaviour. However, this initialization has a better score for the reliability part of the CRPS but shows no improvement for the accuracy scores. We can conclude that to maintain a reliable ensemble, the interval of the uniform distribution must be sufficiently large. Nonetheless, an increase in the interval does not offer much improvement for the trade-off between reliability and accuracy.

4.1.4 Input variable selection

440 The results of the sensitivity scores for each variable are presented in Table 5. Recall that this sensitivity score is the average RMSE of the four reference percentiles and all ensemble members (see Sect. 2.4.6). Therefore, the value of this score is proportional to the importance of a variable for the conversion model. As shown in Table 5, snow depth is, unsurprisingly, the most important input variable.

Figure 7 illustrates how reducing the number of input variables gradually affects the SWE estimation error and the corre-
445 sponding CRPS and ignorance score. This reduction follows Table 5; the least influential input variable (snow density from

Table 5. Ordered results of the sensitivity analysis from the least to most influential variable

Variable	Score
Snow density ERA5	23.2
Number of freeze-thaw cycles	34.6
Average temperature of the last six days	63.9
Number of layers in snow pack	102.4
Total precipitation in the last ten days	102.5
Average age of the snow cover	116.9
Accum. positive degrees since the beginning of winter	124.9
Accum. solid precipitation since the beginning of winter	160.2
Number of days since the beginning of winter	184.6
Total solid precipitation in the last ten days	191.0
Days without snow since the beginning of winter	193.2
Snow depth	972.0

ERA5) is removed first, and so on. Figure 7a and b show that the RMSE, MAE, and CRPS do not increase significantly when the number of input variables decreases to the six most influential variables. A larger increase (worsening) is only observed below this number. This pattern reflects the input variables being dependent on each other as all, except for snow density and snow depth, are derived from the same meteorological data. Therefore, adding more variables does not provide more informa-

450 tion to the model. This result is consistent with the study of Odry et al. (2020). When observing the reliability part of the CRPS (Fig. 7b), the score begins to increase below ten input variables. However, the ignorance score in Fig. 7c increases even when only one variable is eliminated. Thus, more input variables help widen the spread of the ensemble. Nevertheless, the ignorance score tends to flatten as the number of input variables increases. Because we want to preserve the spread of the ensemble and obtain a good portrait of the uncertainty of SWE estimation, we use all 12 variables as inputs in the final set-up.

455 **4.1.5 Determining the number of epochs and the number of hidden neurons for a general MLP model covering Canada**

Figure 8a–d show a general decrease in the MAE, CRPS, ignorance score, RMSE, and the reliability part of the CRPS when the number of neurons in the hidden layer is increased. A greater number of neurons in the hidden layer enables the model to learn more features of the training data set. This entails a better performance of the validation data set. However, there is a

460 threshold beyond which these accuracy values start degrading. This threshold occurs at a smaller number of hidden neurons for a smaller number of epochs. For the MBE in Fig. 8f, we observe no specific trend. Note that the thresholds are earlier for the ignorance score than for the reliability part of the CRPS. Consequently, one can assume that the reliability will also increase for higher numbers of hidden neurons for three, four, and five training epochs.
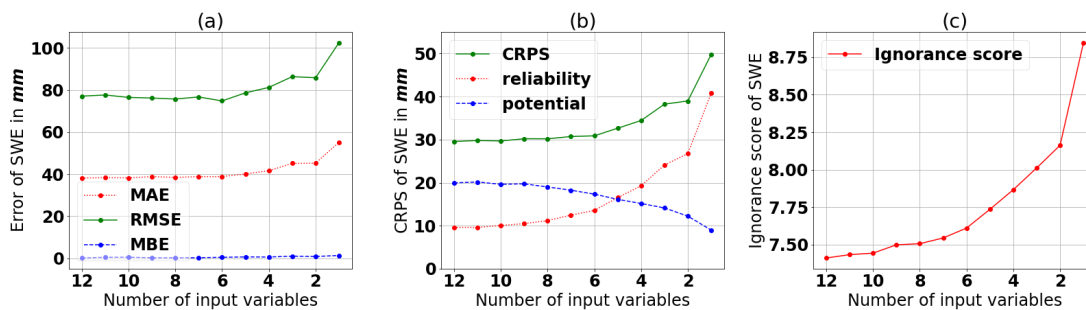
Figure 7. Stepwise reduction of input variables, ordered according to their influence as determined in Table 5
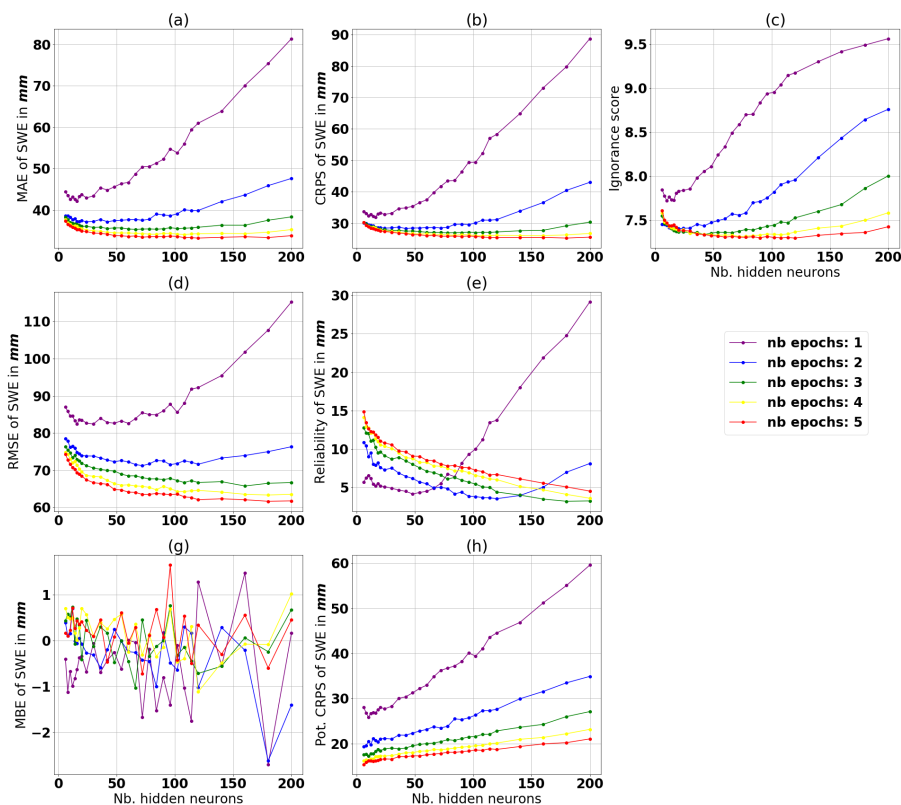
Figure 8. Results of ensembles using the set-up defined in Sect. 4.1 for different numbers of training epochs and neurons in the hidden layer

The earlier increase in the ignorance score can be explained by observing the rank histograms of 1–3 training epochs in Fig.

465 9. A bell shape in the rank histogram indicates that the spread of the ensemble is too wide, resulting in low values overall for the empirical pdf. Because the ignorance score is computed using the logarithm of the probability density corresponding to the observation, overdispersed ensembles strongly penalize this score.
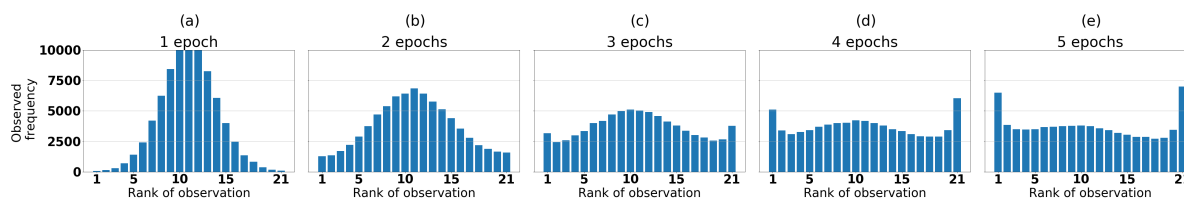
Figure 9. Rank histograms of the MLP ensemble with 120 neurons in the hidden layer for different numbers of epochs

The formula of the reliability part of the CRPS in Eq. (6) demonstrates that the difference between the observed frequency, given the predicted distribution, and the predicted probability is squared. Therefore, if the difference is small, this will dominate the reliability part and result in a small value even with a large ensemble spread. Thus, the reliability score penalizes the outliers in the trials having more than four or five epochs.

This analysis finalizes the first MLP model for Canada as a whole. Our final choice for MLP architecture is the set-up derived from Sect. 4.1; however, we set the number of epochs at five and the number of neurons in the hidden layer at 120. This provides the best trade-off between accuracy and reliability because 120 hidden neurons is the point where accuracy stagnates and where the ignorance score increases for five epochs. Except for the case of one training epoch, the reliability part of the CRPS is poorest for five epochs, but it also leads to an acceptable rank histogram in Fig. 9e. The ensembles trained for two and three epochs show a bell shape, indicating underdispersion, which is an unacceptable condition.

4.1.6 Determining the number of epochs and the number of hidden neurons for a specific MLP model for each snow class

We repeat the same procedure as in Sect. 4.1.5; this time, however, it is run separately for each snow class. Table 6 shows the optimal combination of the number of hidden neurons and number of epochs for each snow class. The behaviour is similar to the "single model" case covering Canada. However, as the amount of available data to train decreases, the number of required training epochs increases. In the present case, the amount of available data varies between the different snow classes, some having only short records. In the end, with the above determined model structure (Sect. 4.1.5), we obtain an ensemble of 20 MLPs for each snow class. This represents the second group of MLP models (one per snow class) that we compare with the single MLP (covering Canada as a whole) and regression models.

4.2 MLP model performance on a testing set

In this section, we use the testing data set and evaluate the performance of the single MLP ensemble model covering Canada as a whole and the separate models for each snow class (multiple MLP ensembles). Table 7 reveals that the multiple MLP ensemble model has a better overall performance. The reliability diagram and the rank histogram also show a more reliable estimate for the multiple MLP ensembles (Fig. 10).

Table 6. Optimal combination of the number of hidden neurons and number of epochs for each snow class

	Ephemeral	Prairie	Tundra	Taiga	Mountain	Maritime
Nb. records (training data set)	1124	3754	4451	7962	22414	38711
Nb. epochs	50	30	30	20	20	20
Nb. neurons	12	24	48	96	192	192

Table 7. Performance evaluation of the model with a single MLP ensemble covering Canada (single MLP) and the multiple MLP ensembles, evaluated using the testing data set

	Single MLP	Multiple MLP
MAE	32.8	29.3
RMSE	61.0	51.5
MBE	0.4	0.6
CRPS	25.1	21.9
Reliability	6.4	3.7
Pot. CRPS	18.6	18.2
Ign. score	7.28	7.22

Figure 11 shows the distribution of the residuals of the simulated ensemble median and the observations. The distribution of the residuals for the multiple MLP model is narrower, indicating less error overall. Both are approximately symmetrical around zero, which strengthens the result of the MBE being almost zero (Table 7). For the single MLP model, 50 % of the errors are smaller than ±20 mm, and 90 % are smaller than ±80 mm. For the multiple MLP model, 50 % of the errors are smaller than ±18 mm, and 90 % are smaller than ±65 mm.

Next, we apply skill scores to ensure a valid comparison between SWE estimates of differing magnitudes. The climatology of SWE from the CHSS data set is used as the reference simulation (see Sect. 2.4.5).

The separation into different snow classes reveals that, as expected, the multiple MLP model eliminates bias in all snow classes, as shown in Fig. 12a. Furthermore, for all snow classes, the two MLP models perform better than climatology, as indicated by positive MAE, RMSE, and CRPS skill scores in Fig. 12a and b.

The model performs better within the *ephemeral snow* class in terms of accuracy, although the results need to be taken with some reservations; the low amount of data and the high variability of SWE in this snow class induce a rather poor reference simulation (climatology).

The reliability part of the CRPS in Fig. 12b shows a large improvement of the multiple MLP model for the *ephemeral snow* class. In contrast, the ignorance score in Fig. 12c shows a slightly poorer skill score for this same snow class. A deeper analysis
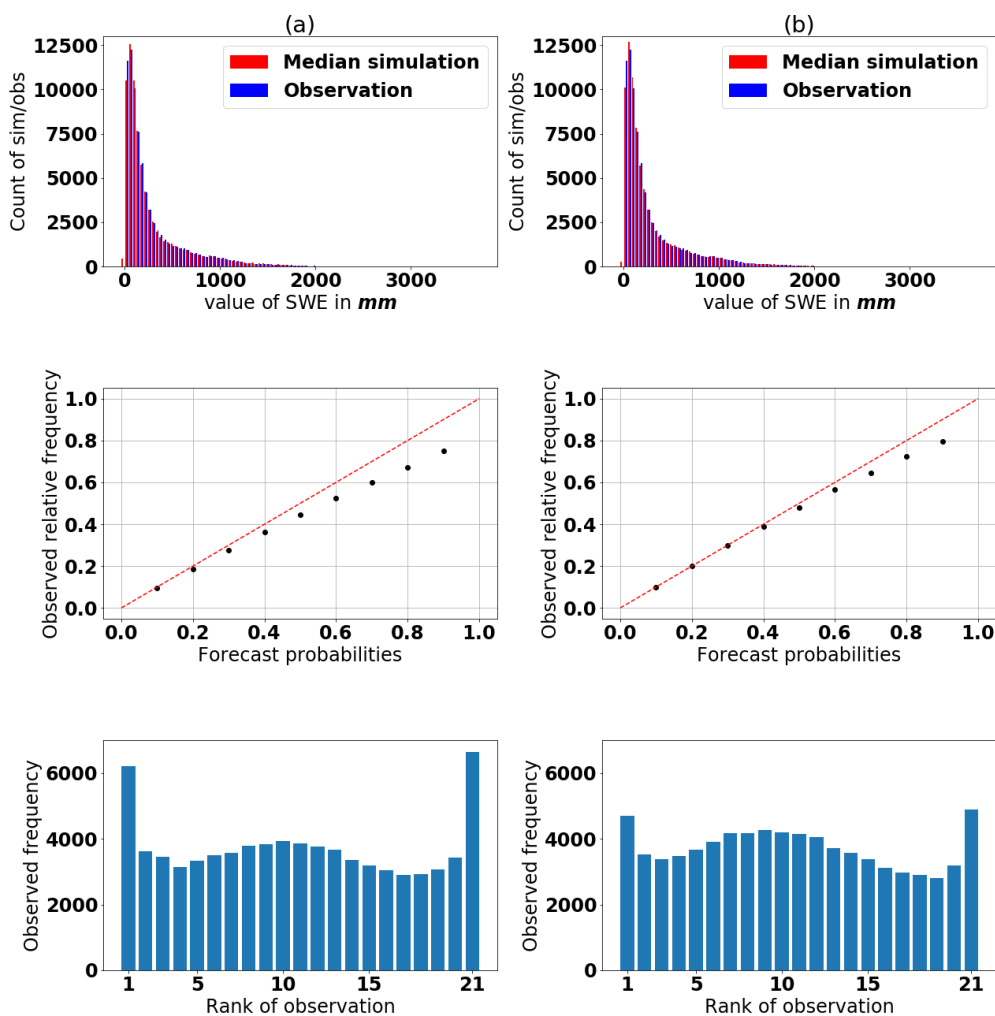
Figure 10. Results for the (a) final single MLP ensemble and (b) the final multiple MLP ensembles, both evaluated using the testing data set

of the *ephemeral snow* class shows that the multiple MLP model has an almost reliable rank histogram, with only a slight tendency toward underdispersion, whereas this underdispersion is more pronounced for the single MLP model. This pattern explains the large improvement in the reliability part of the CRPS. The average spread of the ensemble for the multiple MLP

510 model in this class is twice that of the single MLP model. The larger spread entails lower values in the empirical pdf and therefore lower ignorance scores.
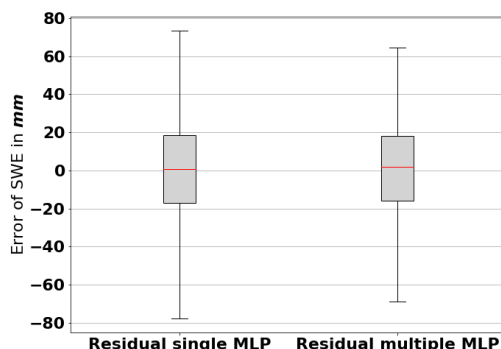
Figure 11. Distribution of the residuals of the simulated ensemble median and the observations. The boundary of the box shows the 1st and 3rd quartiles; the caps at the end of the whiskers show the 5th and 95th percentiles

The *tundra snow* and *taiga snow* classes have the lowest skill scores in terms of both accuracy and reliability. In particular, the reliability part of the CRPS in Fig. 12b results in a large decrease of the skill score. As the variability of SWE is low in these areas, as presented in Fig. 2, the reference simulation based on climatology produces a reliable ensemble and causes a

515   poor reliability skill score for the actual simulation.

The *maritime snow* class shows a slightly better accuracy than the *mountain snow* class. This difference may relate to the complexity of snow accumulation patterns in mountainous regions owing to the high spatial and temporal variability of all physical processes and variables in these areas. Also the temperature correction, as explained in Sect. 3.2, induces larger errors where height is highly variable. The improved reliability skill score for the multiple MLP model is explained by fewer outliers

520   for the *mountain* and *maritime snow* classes. In total, both classes show similar rank histograms for the single and multiple models. The slight decrease in the ignorance skill score for the *maritime snow* class can be explained by different average spreads of the simulated and reference ensembles.

In the next step, the testing data set is divided into elevation classes from 0 m to 2100 m, with a step size of 300 m and one class for sites above 2100 m. The results are not shown, as the main conclusions are identical to those obtained from the

525   separation into snow classes. Both types of MLP ensembles (single and multiple) outperform climatology.

Second, an analysis over the course of the year shows an improved accuracy for the MLP models compared with climatology for all months, except for July, August, and September. During these three months, there is generally very little snow across the country and, consequently, very little data. Additionally, the *beginning of the winter*, subjectively taken as 01 September, causes a reset to zero for several explanatory variables; the variables of temperature, snow depth, and SWE data remain, however,

530   within their usual ranges. This greatly complicates the proper training of the MLPs. Overall, except these three problematic months, the accuracy and reliability remain relatively constant throughout the year with a slightly improved performance in spring and early summer compared with climatology.
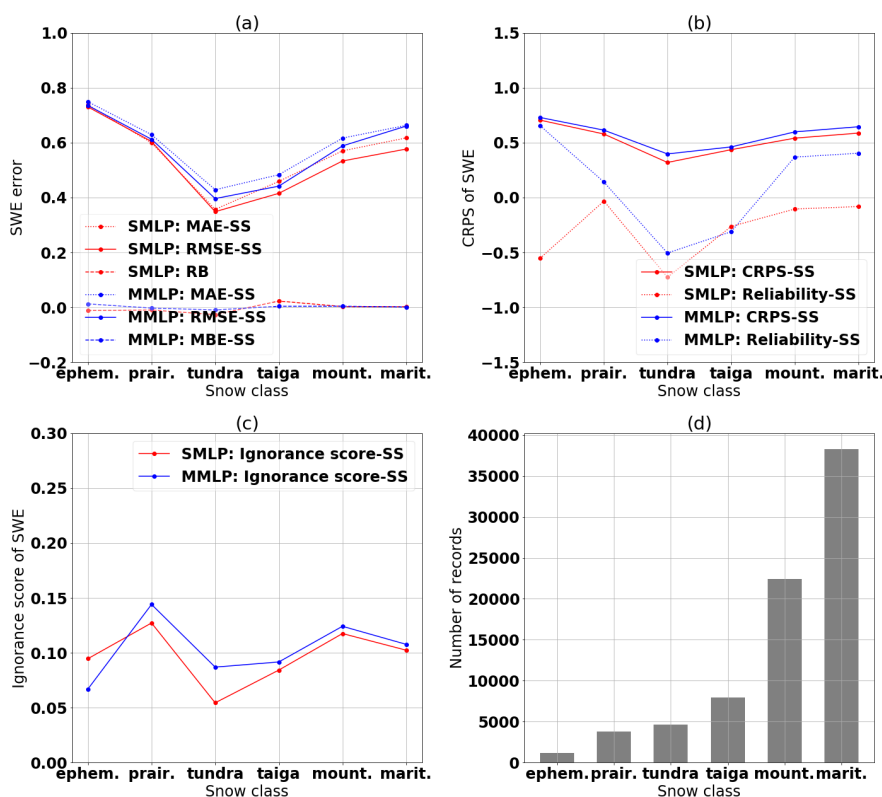
Figure 12. (a)–(c) The performance of the single MLP (SMLP) and multiple MLP (MMLP) models as evaluated for the different snow classes. (d) The number of records in each snow class within the testing data set

4.3 Comparison of MLP models with regression models

The regression models are trained and validated using the same perturbed snow depth data sets that are used to optimize the
535 MLP models in Sect. 3.3. The testing data set is then used for comparison purposes in this section.

There is a possibility of producing an ensemble simulation by performing the deterministic regression models on perturbed snow depth and obtaining multiple members. However, given that the spread in the simulated ensemble of the MLP models is explained mainly by the various parameter initializations, this approach entails an ensemble that is too narrow when regression models are used. This leads to a comparison of models that has little meaning: the optimization of the Sturm model could be
540 initialized with different parameter sets; however, the Jonas model has a perfect set of parameters. Therefore, we must compare the models using exclusively deterministic evaluation techniques.

In addition to the regression models, we also considered a simple benchmark model, which takes the average observed snow density in the training and validation data sets as a constant snow density and then calculates SWE for the testing data set.

Table 8. Comparison of the overall performance evaluation of the two MLP models and the regression models, evaluated using the testing data set

	Benchmark	Jonas	Sturm	SMLP	MMLP
MAE	74.2	44.5	47.7	32.8	29.3
RMSE	145.4	114.7	117.0	61.0	51.5
MBE	-35.3	2.2	-0.5	0.4	0.6
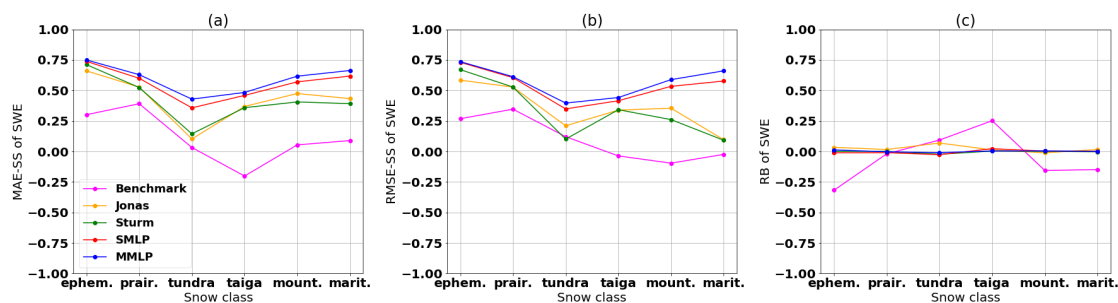
Figure 13. Comparison of the performance of the two MLP and regression models by using (a) the skill score of the MAE, (b) the skill score of the RMSE, and (c) the relative bias (RB) of the SWE simulation, as evaluated for different snow classes

The results are presented in Table 8. Both models using MLP ensembles outperform the simple benchmark as well as the
545   Sturm and Jonas models.

Figure 13 compares the MLP models with the regressions and the simple benchmark for each snow class. The two MLP models outperform both the regressions and the benchmark model for all snow classes. The Sturm model eliminates the bias, as depicted by the relative bias (RB) being equal to zero in Fig. 13c. This removal of bias occurs because the model uses an individual regression model for each snow class. In Fig. 13a and b, the MAE and RMSE skill scores of the Sturm and Jonas
550   models follow mainly the skill scores of both MLP models. However, the Sturm and Jonas models show a decrease in the RMSE-SS for the *mountain snow* and *maritime snow* classes, which indicates a greater number of outliers. One can derive that a single regression model is not sufficient for the high variability in the mountainous region or the spatially large maritime region, which includes two coasts. The better performance of the Jonas model for the *mountain snow* class reflects this model being built and tested for a mountainous region in the Swiss Alps. The benchmark model performs well in the Prairies; this
555   performance can be explained by the flat topography of the Prairies, which may result in a more stable relationship between snow density and snow depth. The poor performance of the *taiga snow* class for the benchmark model relates to the high positive RB of the benchmark model, as presented in Fig. 13c, and the relatively strong performance of the reference simulation for this snow class.
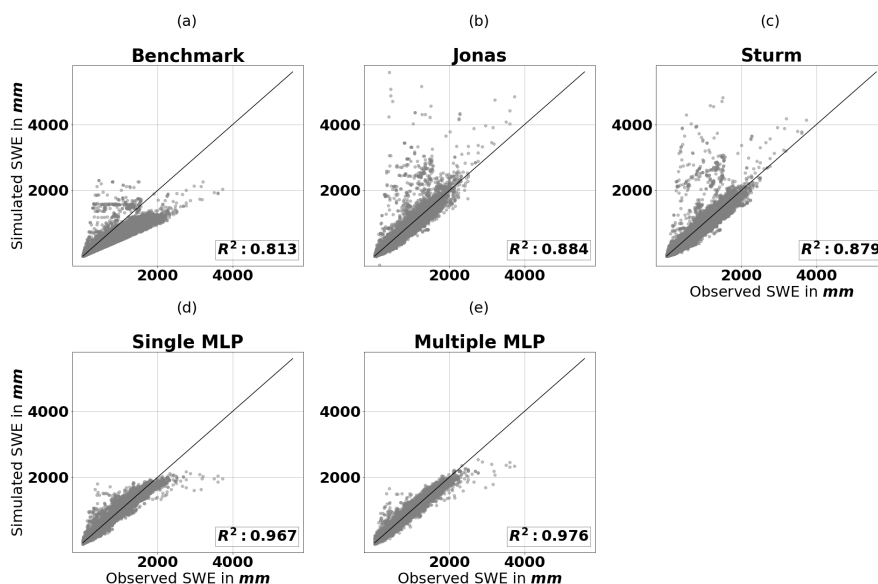
Figure 14. Scatter plot of estimated and observed SWE for each model

Figure 14 shows the simulated SWE against the observed SWE. The benchmark model can explain 81.3% of the variability
in SWE. Figure 14a shows that the model often underestimates SWE, which also explains the negative bias observed in Table 8.
The Jonas and Sturm models explain approximately 7 % more SWE variability. Figure 14b and c reveal that outliers with high
values occur for lower- and medium-range SWE observations. Numerous SWE observations above 2500 mm are consistently
overestimated but follow the identity line. Compared with the two MLP models, the Sturm and Jonas models perform better at
these higher values. Compared with the benchmark model, the single and multiple MLP models explain approximately 15 %
and 16 % more SWE variability, respectively. Figure 14d and e show that up to 2500 mm, the estimates of both MLP models
follow mainly the identity line with a some outliers at approximately 500 mm. Above 2500 mm, the model seems to have an
upper boundary that results in an underestimate. This upper boundary is less restrictive for the multiple MLP model, which
performs better for the high SWE records. The resulting R^2 values for the benchmark and Sturm model are similar to those in
the study of Odry et al. (2020), which considered only the province of Quebec. The MLP model in the latter study achieved an
R^2 value of 0.919. Therefore, we improve on this by 5 % and 6 % for the single and multiple MLP models, respectively.

5   Conclusions

In this study, we test two hypotheses. The first hypothesis holds that using SWE rather than density as the target variable for
the ANN will produce more accurate estimates of SWE. The second hypothesis states that in-depth testing of different ANN
structures will further improve these SWE estimates. As part of this second hypothesis, we investigate whether the ANN model
must be trained specifically for different regions, as determined by snow climate classes (Sturm et al., 2009), or whether the

model could be trained only once, for Canada as a whole. Furthermore, we use existing regression models, developed for the same purpose, as benchmarks to obtain a better perspective on how our model performs.

Our snow-depth-to-SWE model uses the inputs of snow depth, estimated snow density, and other explanatory variables derived from meteorological data. The available snow data includes snow depth, SWE, and snow density measurements from
580 across Canada, collected over almost 40 years.

We then use an ensemble of multiple MLPs to address the issue of the random parameter initialization during optimization. The approach also provides a probabilistic estimate to gain greater insight into model performance. A trade-off between reliability and accuracy is used as a means of evaluation, which gives a more comprehensive analysis of SWE-estimation models.

585 Many previous models (e.g. Odry et al., 2020; Jonas et al., 2009; Sturm et al., 2010; Painter et al., 2016; Broxton et al., 2019) determine SWE by estimating snow density and calculating SWE on the basis of snow density and snow depth. This study investigates a direct estimate of SWE. This approach shows a slight increase in accuracy and a large gain in reliability compared to indirect estimates. Consequently, the remainder of the study uses SWE as the target variable rather than snow density.

590 In our investigation of model structures, we built two models. One model uses a single MLP ensemble for all of Canada. The second model uses additional geophysical information and trains one MLP ensemble for each snow class, as defined by Sturm et al. (2009). Model evaluation of the independent testing data set indicates that the multiple ensemble model outperforms the single ensemble model. Both models show weak performances for high values of SWE. Furthermore, both MLP models outperform existing regression models and a benchmark model based on climatology, and they improve on the basic MLP
595 ensemble model proposed earlier by Odry et al. (2020).

A sensitivity analysis reveals that a greater number of input variables increases the reliability of the ensemble. Therefore, adding more variables could further heighten the model's reliability. An incorporation of these variables (of a type other than meteorological ones) may also help to increase accuracy. Finally, this study shows an optimal performance of networks having large numbers of neurons in the hidden layer at amounts far above the commonly used rules of thumb. This provides a moti-
600 vation to look into network structures having multiple layers. Montúfar (2014) also showed that the number of neurons needed to approximate a function in a deep network with multiple layers increases exponentially in a network with one hidden layer. Therefore, deep networks are possibly more efficient. Goodfellow et al. (2016) provides multiple examples that foster this claim empirically. As a result, there could remain numerous means of improving SWE estimates from snow depth using machine learning techniques, and the methods proposed here could be refined. It could also be useful to investigate the application of
605 other types of machine-learning algorithms, including random forests.

Code and data availability. The code including some testing data is available on Ntokas (2020). The whole data set excluding data from partners which do not allow to publish them (affecting 2%) is available on *Harvard Dataverse*, which link can be found in the ReadMe file on Ntokas (2020). Note that the user needs to fill in a form regarding the terms and conditions for the usage.

Author contributions. Konstantin Ntokas (K.N.) performed all the computations, suggested most of the specific tests to be undertaken, and
610 wrote the initial version of the manuscript. Jean Odry suggested edits to the manuscript and provided the initial version of the codes to K.N.,
which K.N. then translated to Python and modified. The original idea for this work was from Marie-Amélie Boucher, who also guided the
work throughout and edited the manuscript significantly. Camille Garnaud was involved in the guidance for this project as a representative
of ECCC and provided a final proofread of the manuscript before its submission.

Competing interests. The authors declare no competing interests.

615 *Acknowledgements.* The authors acknowledge the financial support of Environment and Climate Change Canada and are also thankful to the
Réseau Météorologique Coopératif du Québec for providing the data required for this study.

References

Anderson, J. L.: A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations, Journal of Climate, 9, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2, 1996.

620 Barnett, T. P., Adam, J., and Lettenmaier, D.: Potential impacts of a warming climate on water availability in snow-dominated regions, Nature, 438, 303–309, https://doi.org/10.1038/nature04141, 2005.

Barry, R. G. and Chorley, R. J.: Atmosphere, Weather and Climate, Methuen, 5 edn., 1987.

Bernier, N. B., Bélair, S., and Bilodeau, B.: Near-Surface and Land Surface Forecast System of the Vancouver 2010 Winter Olympic and Paralympic Games, Journal of Hydrometeorology, 12, 508–530, https://doi.org/10.1175/2011JHM1250.1, 2011.

625 Bertalanffy, v. L.: General system theory : foundations, development, applications, New York : George Braziller, 1968.

Bormann, K. J., Westra, S., Evans, J. P., and McCabe, M. F.: Spatial and temporal variability in seasonal snow density, Journal of Hydrology, 484, 63 – 73, https://doi.org/10.1016/j.jhydrol.2013.01.032, 2013.

Boucher, M.-A., Laliberté, J.-P., and Anctil, F.: An experiment on the evolution of an ensemble of neural networks for streamflow forecasting, Hydrology and Earth System Sciences, 14, 603–612, https://doi.org/10.5194/hess-14-603-2010, https://www.hydrol-earth-syst-sci.net/14/
630 603/2010/, 2010.

Brown, R. D., Fang, B., and Mudryk, L.: Update of Canadian Historical Snow Survey Data and Analysis of Snow Water Equivalent Trends, 1967–2016, Atmosphere-Ocean, 57, 149–156, https://doi.org/10.1080/07055900.2019.1598843, https://doi.org/10.1080/07055900.2019.1598843, 2019.

Broxton, P. D., van Leeuwen, W. J. D., and Biederman, J. A.: Improving Snow Water Equivalent Maps With Machine Learning of Snow
635 Survey and Lidar Measurements, Water Resources Research, 55, 3739–3757, https://doi.org/10.1029/2018WR024146, 2019.

Bühler, Y., Adams, M., Bösch, R., and Stoffel, A.: Mapping snow depth in alpine terrain with unmanned aerial systems (UASs): Potential and limitations, The Cryosphere, 10, 1075–1088, https://doi.org/10.5194/tc-10-1075-2016, 2016.

Casson, D. R., Werner, M., Weerts, A., and Solomatine, D.: Global re-analysis datasets to improve hydrological assessment and snow water equivalent estimation in a sub-Arctic watershed, Hydrology and Earth System Sciences, 22, 4685–4697, https://doi.org/10.5194/hess-22-
640 4685-2018, 2018.

Copernicus Climate Change Service (C3S): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store (CDS), date of access: Oct 2019, https://cds.climate.copernicus.eu/cdsapp#!/home, 2017.

Cybenko, G.: Approximation by superpositions of a sigmoidal function, Math. Control Signal Systems, 2, 303–314, https://doi.org/10.1007/BF02551274, 1989.

645 Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, in: Advances in Neural Information Processing Systems 27, edited by Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., pp. 2933–2941, Curran Associates, Inc., 2014.

Dickinson, W. T. and Whiteley, H. R.: A SAMPLING SCHEME FOR SHALLOW SNOWPACKS, Hydrological Sciences Bulletin, 17, 247–258, https://doi.org/10.1080/02626667209493832, 1972.

650 Dodson, R. and Marks, D.: Daily air temperature interpolated at high spatial resolution over a large mountainous region, Climate Research, 8, 1–20, https://doi.org/10.3354/cr008001, 1997.

Dozier, J.: Mountain hydrology, snow color, and the fourth paradigm, Eos, Transactions American Geophysical Union, 92, 373–374, https://doi.org/10.1029/2011EO430001, 2011.

ECCC Environment and Climate Change Canada: Canadian historical snow survey data, date of access: Oct 2019, http://data.ec.gc.ca/data/
climate/systems/canadian-historical-snow-survey-data/, 2019.

Garnaud, C., Bélair, S., Carrera, M. L., Derksen, C., Bilodeau, B., Abrahamowicz, M., Gauthier, N., and Vionnet, V.: Quantifying Snow
Mass Mission Concept Trade-Offs Using an Observing System Simulation Experiment, Journal of Hydrometeorology, 20, 155–173,
https://doi.org/10.1175/JHM-D-17-0241.1, https://doi.org/10.1175/JHM-D-17-0241.1, 2019.

Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, Journal of Machine Learning Research
- Proceedings Track, 9, 249–256, 2010.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association,
102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning - Chapter 5 - 8, MIT Press, http://www.deeplearningbook.org, 2016.

Hamill, T. and Colucci, S.: Verification of Eta–RSM Short-Range Ensemble Forecasts, Monthly Weather Review, 125, 1312–1327,
https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2, 1997.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Monthly Weather Review, 129, 550–560,
https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in
Statistics, Springer New York, New York, 2 edn., https://web.stanford.edu/~hastie/Papers/ESLII.pdf, 2009.

Heaton, J.: Introduction to Neural Networks for Java, Heaton Research, Inc., 2008.

Hecht-Nielsen, R.: Neurocomputing, Addison-Wesley Longman Publishing Co., Inc., USA, 1989.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15,
559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hock, R., Rees, G., Williams, M. W., and Ramirez, E.: Contribution from glaciers and snow cover to runoff from mountains in different
climates, Hydrological Processes, 20, 2089–2090, https://doi.org/10.1002/hyp.6206, 2006.

Hornik, K.: Approximation capabilities of multilayer feedforward networks, Neural Networks, 4, 251 – 257, https://doi.org/10.1016/0893-
6080(91)90009-T, 1991.

Jennings, K. S., Winchell, T. S., Livneh, B., and Molotch, N. P.: Spatial variation of the rain–snow temperature threshold across the Northern
Hemisphere, Nature Communications, 9, https://doi.org/10.1038/s41467-018-03629-7, 2018.

Jonas, T., Marty, C., and Magnusson, J.: Estimating the snow water equivalent from snow depth measurements in the Swiss Alps, Journal of
Hydrology, 378, 161–167, https://doi.org/10.1016/j.jhydrol.2009.09.021, 2009.

Kim, E., Gatebe, C., Hall, D., Newlin, J., Misakonis, A., Elder, K., Marshall, H. P., Hiemstra, C., Brucker, L., De Marco, E., Crawford, C.,
Kang, D. H., and Entin, J.: NASA's snowex campaign: Observing seasonal snow in a forested environment, in: 2017 IEEE International
Geoscience and Remote Sensing Symposium (IGARSS), pp. 1388–1390, https://doi.org/10.1109/IGARSS.2017.8127222, 2017.

Kinar, N. J. and Pomeroy, J. W.: Measurement of the physical properties of the snowpack, Reviews of Geophysics, 53, 481–544,
https://doi.org/10.1002/2015RG000481, 2015.

Kirnbauer, R., Blöschl, G., and Gutknecht, D.: Entering the Era of Distributed Snow Models: Paper presented at EGS XVIII General Assem-
bly (Wiesbaden, Germany – May 1993), Hydrology Research, 25, 1–24, 1994.

Kwak, Y., Hwang, J., and Yoo, C.-J.: A new damping strategy of Levenberg-Marquardt algorithm for Multilayer Perceptrons, Neural Network
World, 21, 327–340, https://doi.org/10.14311/NNW.2011.21.020, 2011.

Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., and Wood, E. F.: Inroads of remote sensing into hydrologic science during the WRR era, Water Resources Research, 51, 7309–7342, https://doi.org/10.1002/2015WR017616, 2015.

MELCC Ministère de l'Environnement et de la Lutte contre les Changements Climatiques: Données Du Réseau de Surveillance Du Climat Du Québec, québec: Direction générale du suivi de l'état de l'environnement, 2019.

695 Montúfar, G. F.: Universal Approximation Depth and Errors of Narrow Belief Networks with Discrete Units, Neural Computation, 26, 1386–1407, https://doi.org/10.1162/NECO_a_00601, 2014.

Ntokas, K.: konstntokas/Hydrology_ANN_SD2SWE: Publication HESS, https://doi.org/10.5281/zenodo.4276414, https://doi.org/10.5281/zenodo.4276414, 2020.

Odry, J., Boucher, M. A., Cantet, P., Lachance-Cloutier, S., Turcotte, R., and St-Louis, P. Y.: Using artificial neural networks to estimate
700 snow water equivalent from snow depth, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 0, 1–17, https://doi.org/10.1080/07011784.2020.1796817, https://doi.org/10.1080/07011784.2020.1796817, 2020.

Olden, J. D. and Jackson, D. A.: Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks, Ecological Modelling, 154, 135–150, https://doi.org/10.1016/S0304-3800(02)00064-9, 2002.

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., Hedrick, A., Joyce, M., Laidlaw, R., Marks, D.,
705 Mattmann, C., McGurk, B., Ramirez, P., Richardson, M., Skiles, S. M., Seidel, F. C., and Winstral, A.: The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo, Remote Sensing of Environment, 184, 139 – 152, https://doi.org/10.1016/j.rse.2016.06.018, 2016.

Pomeroy, J., Pietroniro, A., Fang, X., Shaw, D., Armstrong, R., Shook, K., Comeau, L., Toth, B., and Martz, L.: Canadian prairie drought hydrology, p. 59–62, Stewart R, Lawford R (eds) Drought Research Initiative, 2011.

710 Roulston, M. and Smith, L.: Evaluating Probability Forecasts Using Information Theory, Monthly Weather Review, 130, 1653–1160, https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2, 2002.

Schaul, T., Antonoglou, I., and Silver, D.: Unit Tests for Stochastic Optimization, International Conference on Learning Representations (ICLR), https://arxiv.org/abs/1312.6055, 2014.

Seibert, J., Jeníček, M., Huss, M., and Ewen, T.: Snow and Ice in the Hydrosphere, in: Snow and Ice-Related Hazards, Risks, and Disasters,
715 chap. 4, pp. 99–137, Academic Press, https://doi.org/10.1016/B978-0-12-394849-6.00004-4, 2014.

Sturm, M., Holmgren, J., and Liston, G. E.: Global Seasonal Snow Classification System, date of access: Dec 2020, https://doi.org/10.5065/D69G5JX5, 2009.

Sturm, M., Taras, B., Liston, G. E., Derksen, C., Jonas, T., and Lea, J.: Estimating Snow Water Equivalent Using Snow Depth Data and Climate Classes, Journal of Hydrometeorology, 11, 1380–1394, https://doi.org/10.1175/2010JHM1202.1, 2010.

720 Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Workshop on Predictability, 20-22 October 1997, pp. 1–26, ECMWF, ECMWF, Shinfield Park, Reading, https://www.ecmwf.int/node/12555, 1997.

WMO: Guide to Meteorological Instruments and Methods of Observation, 2014 edition, Updated in 2017, https://www.wmo.int/pages/prog/www/IMOP/CIMO-Guide.html, 2014.

Zeiler, M. D.: ADADELTA: An Adaptive Learning Rate Method, https://arxiv.org/abs/1212.5701, 2012.