Estimating SWE in North America and Exploring the Generalization Error Using XGBoost and Random Forest

By

Eirik Storrud Røsvik

Master Thesis
GEO5960
60 credits

Department of Geosciences
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

05/2022

Sammendrag

Estimering av snøens vannekvivalente dybde har lenge vært et mål. Snøsmelting er en viktig kilde til ferskvann for mange regioner i kalde klimaer og utnyttes som en verdifull ressurs innen jordbruk, kraftproduksjon og som drikkevann. Målinger av snøens vannekvivalent er kostbare og ressurskrevende, noe som har skapt et behov for metoder for å estimere dem. I denne artikkelen presenteres fem forskjellige metoder for å gjøre dette. Data hentet fra USA og Canada, bestående av 314105 målepunkter, danner grunnlaget for modellene. I tillegg har variabler blitt hentet fra ERA5-Land og prosessert. En block bootstrap metode har blitt anvendt for å undersøke generaliseringsfeilene i modellene. XGBoost og random forest er to nye metoder som har blitt undersøkt. I tillegg har en multilayer perceptron modell, foreslått av Ntokas et al. (2021), og to regresjonsmodeller foreslått av Jonas et al. (2009) og Sturm et al. (2010) blitt konstruert. Regresjonsmodellene presterte dårligst, mens en R^2 score på over 0.98 ble oppnådd for de øvrige modellene. En høy generaliseringsfeil tyder likevel på at disse resultatene stammer fra overtrente modeller, da den høye nøyaktigheten ikke kunne bli reprodusert for nye regioner. Ved å anvende lokale målinger for å validere modellene ble generaliseringsfeilen redusert og modellene viste bedre resultater. XGBoost modellen hadde de beste resultatene med en R^2 score på 0.93, mens både random forest og multilayer perceptron modellen hadde en R^2 score på over 0.89 Random forest modellen utpreget seg ved å ha den laveste generaliseringsfeilen når den ble anvendt på et usett område. Bruk av kun regionale målepunkter viste seg å gi de beste resultatene ved bruk av XGB modellen, på tross av et mye mindre treningsdatasett. Det er blitt vist at reanalysert meteorologiske data bidrar til bedre modeller. For den beste XGBoost modellen anvendt til å estimere vannekvivalent i en nytt område, sto ERA5-Land deriverte variabler for 10% av modellens nøyaktighetsøkning. Forskjellige metoder har blitt foreslått ut i fra datagrunnlaget i området. Dersom ingen data eksisterer er random forest den modellen som presterte best, ettersom den hadde lavest generaliseringsfeil. Dersom noe data eksisterer kan den brukes til å validere XGBoost modellen for å oppnå bedre resultater. Der hvor tilstrekkelig data eksisterer ble det funnet at en XGBoost modell trent på disse punktene vil gi de beste resultatene. En score på 0.94 ble oppnådd ved denne metoden i Alaska.

Acknowledgements

First, I want to express my gratitude to my supervisors Thomas Vikhamar Schuler and Simon Filhol for their help and support in guiding me through this process. Their help in setting the project of on the right course and bringing the project to a conclusion is truly appreciated.

Contents

Abstract

The Snow Water Equivalent (SWE) is a key area of research when aiming to quantify freshwater resources in cold areas. Snow melt is the main source of fresh water in many cold regions and is a valuable resource for agricultural, hydroelectrical, and community water supply uses. The acquisition of SWE measurements is costly and labor intensive, heightening the need for a computational approach. In this paper two decision tree machine learning techniques, XGBoost and random forest, are suggested as methods to estimate SWE using snow depth, elevation and reanalyzed ERA5-Land data as input variables. These are all variables that are relatively easy to acquire on a regional level. The models are built upon 314,105 SWE measurements collected from historical datasets from Canada and the United States. The generalization error has been analyzed with the use of a block bootstrap dataset division. For comparison, three previously suggested models; a multilayer perceptron model by Ntokas *et al.* (2021) and two regression models by Jonas *et al.* (2009) and Sturm *et al.* (2010), have been constructed. R^2 scores greater than 0.98 were obtained from the non-regression models, but high generalization error found that the performance is not transferable in space. R^2 scores greater than 0.89 were obtained for a novel region using a block bootstrap approach. The XGBoost proved the best performing model with an R^2 score of 0.934. The random forest model was however the model with the lowest generalization error. The non-regression models performed much better than the regression models, showing that modern machine learning techniques are well suited for estimating SWE. The multilayer perceptron model was outperformed by the XGBoost and random forest both in accuracy and computational cost. ERA5-Land derived data was found to be important to the model predictions. In the best performing XGBoost model, ERA5-Land data accounted for 10% of the model gain. Different approaches to model SWE have been suggested based upon local data availability. The random forest should be considered if no local data are available, while the XGBoost should be used where data are available. Local measurements were used to improve models when used for validating model parameters, increasing the XGBoost R^2 score from 0.87 to 0.93. The best results were found when training the XGBoost model on local data, improving the R^2 score from 0.93 to 0.94.

1. Introduction

Precipitation falling as snow plays an important role in many of the world's watersheds. In polar and high elevation regions, where snow falls and accumulates, large amounts of fresh water can be stored in the form of snow. Several points of interest arise around the accumulation of snow and the subsequent melting. Snow melt is the largest contributor to the runoff of many rivers in the western United States and Canada with 50-80% of the runoff occurring in the melt season (Stewart *et al*., 2004). This melt is the primary freshwater source in the western United States (Bales *et al.,* 2006). Runoff in these rivers is often collected in reservoirs, as a steady flow of water is often sought after. These water resources can then sustain agricultural irrigation and urban centers throughout the dryer American summers. Before the start of the melt season these reservoirs are prepared and serve as flood protection systems throughout the melt phase (Dettinger and Cayan, 1995).

With global temperatures increasing (Cook *et al*., 2013), the global snow resources are changing and will likely change further as the earth's atmosphere and oceans reach a higher energy equilibrium. (Mote *et al*., 2018). Changes in seasonal snow cover have been documented all over the world, with snow cover extent and snowfall volume decreasing (Essery, R. 1997; Demaria *et al.,* 2016; Mote *et al.,* 2018). In areas with reliable seasonal snow cover, these changes can impact both the local ecosystems (Vincent, W. 2010) and affect human activities (Prowse *et al.,* 2009).

The water management issues that arise in areas where snow and snow melt are dominant factors often require information regarding the volume of water existing in the watershed as snow (Bales *et al.*, 2006). The snow water equivalent (SWE) is the depth of water that a snowpack would produce if it were to completely melt. It is a key variable to obtain when doing hydrological models in areas with snow cover. A module of calculating the SWE is therefore included in several hydrological models (Ntokas *et al.,* 2021). Knowing the water content locked up in snowpacks is of particularly great interest in the fields of hydroelectricity (Magnusson *et al.,* 2020), and flood prevention (Skaugen, T. 1998).

Several approaches have been proposed to model the SWE of a snowpack. Regression models have been constructed using *in-situ* snow variables like snow depth, elevation and time (Jonas *et al.,* 2009; Sturm *et al.,* 2010). Although in recent years more complex machine learning techniques like artificial neural networks have been investigated (Odry *et al.,* 2020, Snauffer *et al.,* 2018; Ntokas *et al.,* 2021). These have been constructed using several explanatory variables from reanalyzed meteorological data. Predicting SWE is a challenge as snow densification is highly localized (Zhong *et al.,* 2021) and several models that have evaluated their findings have found large biases (Xu *et al.,* 2019). The use of reanalyzed meteorological data has become popular with the release of the European Centre for Medium-range Weather Forecasts (ECMWF) ERA datasets (Hersbach *et al.*, 2020). The latest iteration ERA5 and its land-focused and downscaled ERA5-Land version have been used in several well performing hydrological models (Tarek et al, 2022; Ntokas *et al.* 2021).

1.1 Study Goals

This paper seeks to construct five different models to estimate SWE. An XGBoost, a random forest, a multilayer perceptron, and two regression models are studied. To model the SWE, snow depth, location and reanalyzed data are used as variables. The models are trained with data collected from Canada and

the United States going back to 1980. The Canadian Historical Snow Survey (CHSS) and the US Snow Telemetry Network (SNOTEL) was chosen as the measurement foundation. The goal is to produce a model that can generalize reanalyzed weather data coupled with limited *in-situ* data, chiefly snow depth. To evaluate the model findings and explore any generalization error, a portion of the dataset, the Alaskan SNOTEL stations, have been selected for validation.

1.2 Snow Measurements and Variables

Snow measurement stations commonly measure several physical quantities at the site. The relevant variables for undertaking an analysis on SWE are snow depth SD, density ρ and SWE. The depth-averaged snow density $\bar{\rho}$ (g/cm^3) is defined in Equation 1, with m_s being the mass of the snow and V_T the total volume. (Kinar and Pomeroy, 2015) A survey conducted on 37000 measurements from Norway found the snow density ranges from 0.052g/cm^3 to 0.656 g/cm^3 with an average of 0.33 g/cm^3. (Bruland *et al.*, 2015)

$$\bar{\rho} = \frac{m_s}{V_T} \quad \text{Equation 1}$$

$$SWE = SD * \bar{\rho} \quad \text{Equation 2}$$

Snow depth (cm) is manually measured using a specialized measuring stick that is pushed through the snow down to the ground. This has the advantage of being very precise, but the method is labor intensive and can't easily be scaled up (Kinar and Pomeroy, 2015). Several CHSS and SNOTEL stations have ultrasonic snow depth sensors, getting precise depth measurements at frequent intervals (Vionnet *et al.,* 2021, Anderson & Wirt, 2008). These sensors are still susceptible to errors as damage to or blockage of the transducer can give faulty measurements (Anderson & Wirt, 2008). These tools give a good foundation of precise measurements to calibrate remote sensing equipment. Light Detection and Ranging (Lidar) is a remote-sensing technology used to obtain precise altimetry data. Lidar measurements can be obtained using airplanes for larger areas or from ground-based Lidar systems for the immediate surroundings. Snow depth can be calculated from airborne altimetry by comparing the snow free elevation with the snow-covered elevation whereas ground-based systems use distance between known points in order to derive depth (Deems *et al.,* 2013).

SWE is the amount of water contained in the snowpack and is measured in mm spread over 1 m^2. The SWE depth is manually measured using gravimetric data. This can be done through digging a snow pit, measuring the density and thickness of each layer (Fierz et al, 2009). A less laborious method is extracting a cylindrical core of snow which is then weighed. If the snow depth and base area of the cylinder are known, the bulk density is calculated using Equation 01. Measuring stations for SWE often use snow pillows. These are bags filled with an anti-freeze liquid, and as snow falls on top of the pillow the pressure inside is measured to calculate the weight, getting the SWE through Equation 02 (Kinar and Pomeroy, 2015). Johnson (2004) showed that snow pillows are prone to errors and found three main sources: (1) With differences in compressibility between the snow pillow and the surrounding snow, a shear stress along the border of the sensor can occur. (2) When the temperature at the ground level is 0 °C a difference in thermal conductivity can cause uneven melt rates, which again can cause a similar shear stress. For longer periods with uneven melt rates a difference can develop between the snow above the sensor and

surrounding snow. (3) The last identified source of error are sensors with high compressibility. For higher SWE levels the compressibility can cause shear stress along the perimeter if the overlying snow sinks unevenly as the snow pillow is compressed (Johnson, 2004). Snow pillow errors are tested through removing all the snow directly above and measuring the weight again, comparing it to that of the pillow (Davis, 1973).

SWE measurements are also hard to do remotely due to the heterogeneous density distribution of snow.

Since a snowpack evolves throughout the season and the bulk density generally increases with time, the age of the snow can give valuable information. One commonly used parameter is the day of year. September 1st is often used as the start of the snow season in the northern hemisphere.

Precipitation measuring stations in colder climates that experience regular snowfall are designed to collect both solid and liquid precipitation. When snow falls into the measuring bucket the reading will be the water equivalent depth (mm) as the measuring stations ensure that snow melts inside the bucket. This is done with a chemical solution that lowers the freezing point (Rasmussen *et al.,* 2012). In order to predict what precipitation has fallen in solid or liquid form, Jennings *et al.* (2018) proposed an equation to estimate the probability of precipitation falling in solid form.

$$p(T_{av}) = \frac{1}{1+e^{-1.54+1.24T_{av}}} \qquad \text{Equation 3}$$

Where $p(T_{av})$ is the probability of solid precipitation and T_{av} is the average temperature (Jennings *et al.,* 2018).

Zhong *et al.* (2021) found that the distribution of snow properties, including density, is highly localized and varies based on local features like terrain and vegetation. Other local factors that influence the snow depth/SWE relationship include the surrounding vegetation density and type, shading and terrain aspect. Wind induced deposition of snow is also very localized and has been found to result in denser snow.

2. Literature

In this section background knowledge for the most essential components to the research conducted is presented. Section 2.1 describes the snow classes, while section 2.2 covers the reanalyzed data. Section 2.3 to 2.6 covers machine learning and previous machine learning approaches to estimating SWE. Section 2.7 and 2.8 covers the theory regarding the new proposed approaches random forest and XGBoost respectively. Section 2.9 and 2.10 give an overview in ways to evaluate model predictions.

2.1 Snow classes

Sturm and Holmgren developed in 1995 a snow classification system. The six classifications differentiate areas where snow density develops differently. The classifications aim to differentiate areas where different processes like wind, rain, or vegetation dominate the snow densification. This classification system has been used in several SWE models, leading to increased accuracy for models trained independently for each snow class (Sturm *et al.,* 1995). In 2021 the classifications were revised and updated using ERA5-Land to increase the resolution to 10 by 10 arc seconds. Some of the class names were

also changed, as Taiga was renamed to Boreal Forest and Alpine to Montane Forest (Sturm *et al.,* 2021). The snow classes as described in the 1995 paper are:

- Tundra: Cold areas with thin snow cover (<75 cm) and frequent wind. Areas are typically found above the tree line and/or far north.
- Boreal Forest: Cold areas with moderate snow depths (<120 cm). Found in forests where wind and snow density are low.
- Montane Forest: Intermediate to cold deep snow cover (<250 cm). Areas are dominated by low density snowfall with occurring but insignificant melt features. Includes both sub alpine forests and montane areas.
- Maritime: Areas with warm deep snow cover with max snow depth in excess of 300 cm. Snow in these areas have frequent melt features, with coarse grained snow being ubiquitous.
- Prairie: Areas with thin snow cover (<100 cm). Areas are dominated by wind drift and wind slabs.
- Ephemeral: Areas with extremely thin snow cover (<50 cm). In these areas snow starts melting shortly after snowfall, usually melting away before the next snowfall.

These snow classes have been used in several SWE models as either a parameter or as a division when training multiple models. (Jonas *et al*., 2009; Bruland *et al.,* 2015; Ntokas *et al.,* 2021)

2.2 ERA5-Land

The ERA5 dataset is the fifth generation ECMWF atmospheric reanalysis, covering the period from 1950 until the present. The ERA5-Land is another dataset created by ECMWF, covering the period from 1950 until 2-3 months before present. It is created with high resolution numerical integrations of the ECMWF land surface model, using downscaled meteorological forcing from ERA5. ERA5-Land focuses on describing the water and energy cycles on land and is created specifically with hydrologic modeling in mind. The primary advantages in using the ERA5-Land dataset compared to ERA5 is the increased grid cell resolution to 9 km (ERA5-Land) compared to 31 km (ERA5). The disadvantage to using the ERA5-Land dataset is the delay in data availability. The up to 3-month delay makes real time analysis with ERA5-Land impossible. (Sabater *et al.,* 2021)

ERA5 and its predecessors have been widely used in climate models since their release. The benefits of being able to rely on reanalyzed data instead of actual observation are many. For hydrological models, observations in some regions are too scarce to build models upon, and areas where one parameter has been measured others may be missing. It is therefore common to use reanalyzed data as meteorological forcing when working with region-sized climate models. (Tarek et al, 2020)

2.3 Machine learning

Machine learning is a term coined in 1959 when computer software was used to play simple strategic games with performances surpassing that of the human who made it. Machine learning algorithms, and especially Artificial Intelligence (AI), are probabilistic and iterative methods that require large amounts of input data and are notoriously known for their computational cost. Due to these limitations, models using AI before the year 2000 would incorporate a large amount of human expertise in their models. These expert systems use AI in order to approximate a human expert's assessment of the input variables.

Increased access to computing power coupled with greater availability of digitized data after the advent of the Internet has led machine learning to become relevant again (Lange and Sippel, 2020).

2.4 The Sturm Model

Sturm *et al.* (2010) developed a regression model for estimating snow bulk density using snow depth, day of year, and snow class as input parameters. The model is trained for each snow class, meaning the dataset is divided into subsets containing each snow class. The regression model therefore has a unique set of parameters for each class.

$$\rho_{sim} = (\rho_{max} - \rho_0)[1 - e^{-k_1 SD_{obs} - k_2 DOY_{obs}}] + \rho_0 \qquad \text{Equation 4}$$

Where ρ_{max}, k_1, k_2, and ρ_0 are the parameters tuned for each snow class. ρ_0 is the initial density of each individual snow layer and ρ_{max} is the maximum density in each snow class subset. The fitting parameters k_1 and k_2 are the densification factors of snow depth and day of year respectively.

To quantify the model fit, the root mean square error between the estimated and observed snow densities is used. (Sturm *et al.,* 2010)

2.5 The Jonas Model

Jonas *et al.* (2009) proposed a linear regression model to estimate the snow's bulk density. The Jonas model uses four parameters that were found to have a notable effect on snow densification. The parameters used are the season, snow depth, elevation, and region. The Jonas model splits the dataset into several subsets depending on altitude and season of the measurements. They propose three divisions of the dataset based on elevation, (1) <1400m, (2) ≥1400m and <2000m, and (3) ≥2000m. Twelve temporal divisions are proposed, each containing one month. The linear regression model uses the equation

$$\rho_{sim} = a\, SD_{obs} + b + offset_{reg} \qquad \text{Equation 5}$$

Where *a* and *b* are the parameters optimized by the regression and $offset_{reg}$ is the regional specific parameter. For each region the averaged predicted densities are subtracted from the averaged measured densities to find the offset value. The Jonas model was conducted in the Swiss alps and the regional divisions were done based on geographic features. (Jonas *et al.,* 2009)

For this paper, the regions used are the snow classes defined by Sturm *et al.* (1995) as proposed by Ntokas *et al.* (2021). The regional offset is then calculated independently for each elevation and month to reduce regional bias.

2.6 The Ntokas Model

Ntokas *et al.* (2021) proposed an ensemble multilayer perceptron (MLP) model to calculate SWE directly using snow depth, elevation, snow classes and ERA5-Land data. Ntokas *et al.* argue for modeling SWE directly instead of bulk density which is later converted to SWE. Their model was produced using the CHSS dataset that was randomly split into three parts. One part was used for training the model, one part was used for validation and finding the best parameters, while the last third was used to test the model findings

with unseen data. The setup for the ensemble MLP model was found using the ceteris paribus principle where one parameter is fitted while keeping the rest static. The best parameters were found with all snow classes combined will be described in section 3.3. For the final model the snow classes were split up into separate datasets and the number of hidden layers and epochs were tested for each one. The MLP model is constructed of 20 ensemble members that are all trained and the final predicted value is the median of the member predictions.

The ERA5-Land data used as input parameters gives the Ntokas model a much higher degree of complexity compared to the previous Jonas and Sturm regression models. While it has higher complexity, the *in-situ* measurements needed remain the same. The ERA5-Land data used is temperature, precipitation and snow density. Daily T_{min}, T_{max}, and precipitation is used to calculate further explanatory variables used in the model. These variables come in addition to the *in-situ* measurements. All the explanatory variables used in the Ntokas MLP model are based on Odry *et al.* (2020) except for ERA5 snow density.

Elevation of the observations is used in order to correct the temperature data from ERA5-Land, since the station elevation can differ significantly from the ERA5-Land grid cell elevation. The correction equation used is:

$$T_{cor} = T_{ERA5} + lapserate\left(\frac{Elev_{obs} - Elev_{ERA5}}{1000}\right) \quad \text{Equation 6}$$

Where T_{cor} is the corrected temperature, T_{ERA5} is the temperature from the ERA5-Land dataset, and $Elev_{obs}$ and $Elev_{ERA5}$ is the elevation of the observation and ERA5-Land grid cell respectively. The lapse rate used by Ntokas *et al.* (2021) is $-6^{o}Ckm^{-1}$.

Ntokas *et al.* (2021) concluded that predicting SWE directly yielded better model performance than when predicting density later converted to SWE. They produced two versions of the model, one where all snow classes were combined, and a second model that contains six MLPs individually trained for each snow class. Of these two the latter version had the best performance.

2.7 Random forest algorithms

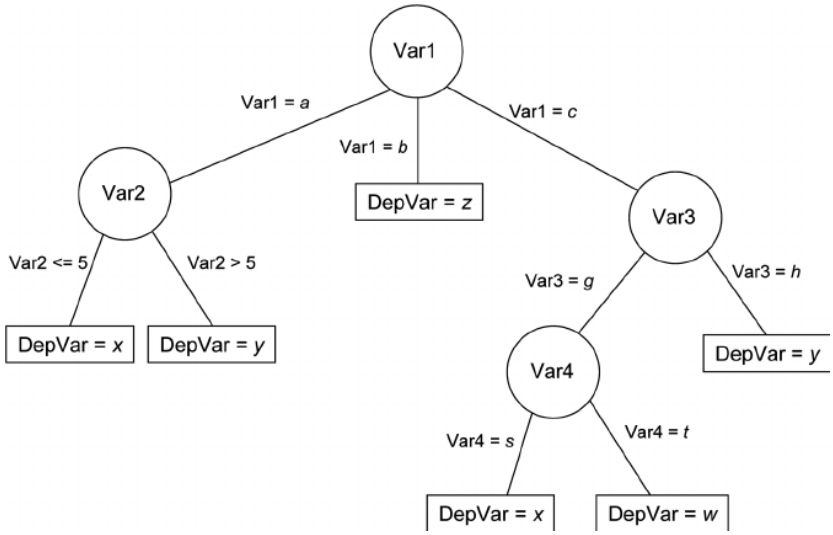

Figure 2.1: Example of data flow in a CART, nodes (circles) decide what variable is tested while the branches (lines) represent the test being performed. The leaves (squares) represent the final prediction of the tree. (Murray and Scime, 2010)

The enhanced classification and regression tree (CART) method random forest (RF) was developed by Breiman in 2001. RF algorithms are designed around the concept of decision trees (Figure 2.1). CARTs are flowchart-like structures where the input data are processed. Each split represents a test of the input data, and each branch leads to a new test depending on the outcome. These splits are formed using a randomly selected set of variables that the tree is constructed with. From this set of available predictors, the best split is chosen. At the end of the tests is a leaf, which will be the result of the input data. The final output of the RF algorithm is the average value of the ensemble CART outputs. RF is considered an ensemble learning technique as many decision trees are constructed in order to reduce variance. This also results in RF models being black-box algorithms, as it is impossible for a human to understand all the trees. (Breiman, 2001)

There are several hyperparameters that need to be tuned when constructing an RF. The most important parameters to be determined are the number layers (layers) and the number of predictors tested at each node (max features). These are optimized to reduce both the generalization error and correlation among CARTs. The way decision tree models only evaluates one feature at a time makes them immune to poor performance stemming from variable collinearity.

The number of trees (estimators) used in the ensemble can be decided. While not a tuning parameter it is recommended to use as many trees as possible to ensure that each candidate feature has the opportunity to be selected. The max feature parameter decides the number of features randomly selected as candidates when constructing each tree. A lower number of candidates can lead to increased precision if there are other features with a big impact on the result, while a higher number will reduce the risk of having non-informative candidate features (Liaw & Wiener, 2001). The default value for how many

variables used for constructing the RF regressor in the Python scikit-learn package is the square root of the total number of variables.

2.8 Extreme Gradient Boosting (XGBoost)

The gradient tree boosting algorithm XGBoost was developed by Chen and Guestrin with a scalable implementation released in 2016. XGBoost is a preferred tool amongst programmers developing machine learning models, being the most used algorithm in winning submissions to Kaggle competitions in 2015. This is due to XGBoost being one of the best performing algorithms coupled with its high execution speed (Chen and Guestrin, 2016).

XGBoost is a CART model, but unlike the random forest it uses gradient boosting. When optimizing the predictions of the target variable, gradient boosted algorithms will use a combination of several weaker learners. For regression problems the XGBoost model uses regression trees as the weak learners. These weaker regression trees follow the same general structure as the decision trees described in section 2.7. XGBoost differs in that it minimizes a regularized objective function. This combines a convex loss function calculated from the residual errors of the prediction and a second step that penalizes model complexity. Trees are added to the model iteratively with the next tree predicting the residuals of the prior trees. All the constructed trees are in the end used to form the final model prediction. The model gets is name from the way the loss is minimized when adding new trees, depicted in Figure 2.2. (ASM, 2021)

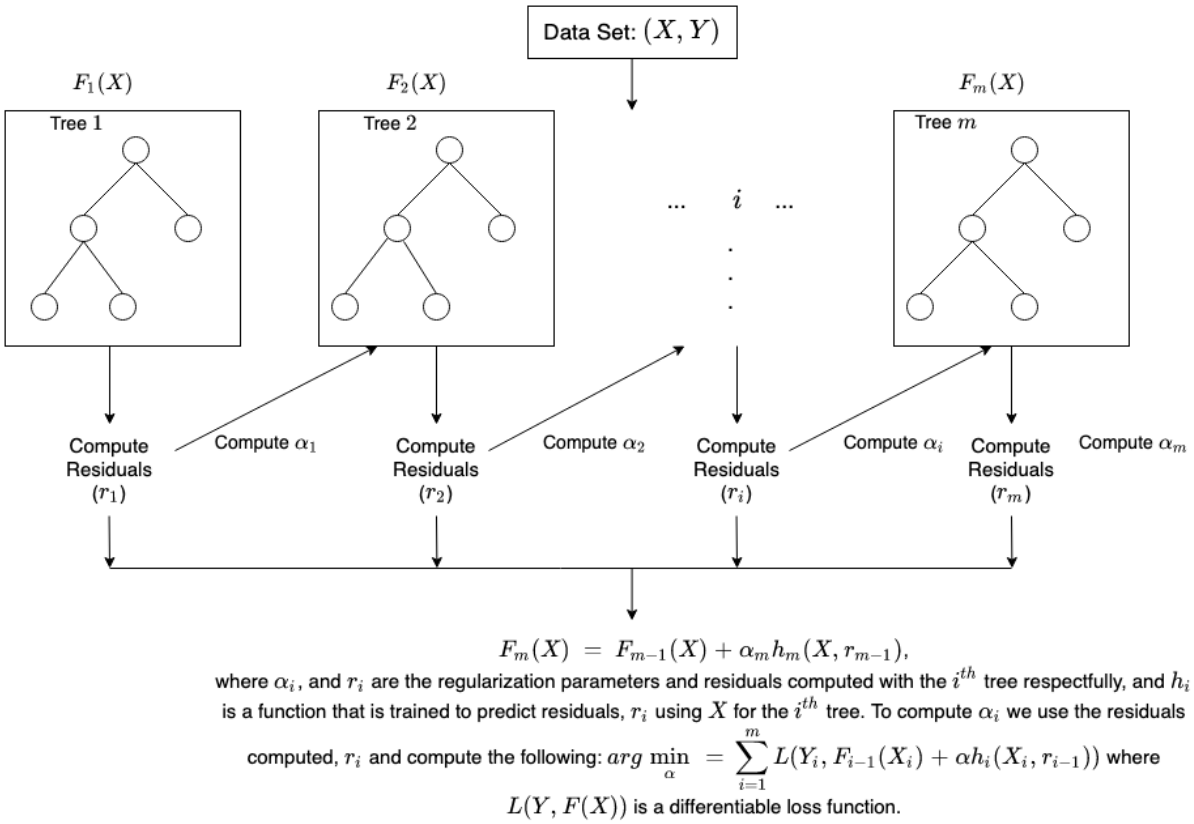

Figure 2.2: Depiction of XGBoost model construction from Amazon SageMaker Developer Guide (ASM, 2021).

2.9 Model Evaluation

Several measures can be used to evaluate the model performance. These methods compare the model predictions with actual SWE measurements. The data points used for measuring the performance of the model are not used to train the model. The general goal for any model is to reduce the difference between predicted and true values, and several different metrics exist to measure this. The most widely used metrics are Mean Absolute Error (MAE), the Mean Squared Error (MSE) and its root (RMSE), and the coefficient of determination (R^2) (Chicco *et al.,* 2021). These deterministic metrics can tell a great deal about the model performance, but other metrics have also been used to get further insight. Probability density functions (PDF) obtained from a trained model can also be used to evaluate its accuracy. A PDF is in this paper defined as a vector where the *i*-th component of the vector describes the probability of the *i*-th outcome occurring so that the sum of all components equals 1 following Roulston and Smith (2002).

The most common benchmarks for evaluating a model's performance are MAE, MSE, RMSE, R^2, and the mean bias error (MBE). These algorithms take as input the predicted $\hat{y}$ and the true y. For the MLP ensemble model, the median of the predicted values is applied.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{Equation 7}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{Equation 8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{Equation 9}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \qquad \text{Equation 10}$$

The MBE is not an evaluation of model precision but rather the average value of the errors. The bias in the model is its tendency to under- or overestimate over all predicted values (Willmott and Matsuura, 2005).

The coefficient of determination is calculated using the mean $\bar{Y}$ of the true values

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{Equation 11}$$

And the mean sum of squares (MST)

$$MST = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2 \quad \text{Equation 12}$$

$$R^2 = 1 - \frac{MSE}{MST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{Y} - y_i)^2} \quad \text{Equation 13}$$

These metrics differ in how they penalize outliers. MAE does not penalize outliers too much and could be the better metric if the outliers in the dataset stems from corrupted or erroneous measurements. MSE and RMSE, which is monotonically related through the square root, will on the other hand penalize outliers harshly and is a good metric should the model performance on outliers be of importance. Since MST if fixed for a given dataset, R^2 is negatively and monotonically related to MSE (Chicco et al, 2021).

When model performance is compared in this paper, unless otherwise specified, it is the respective R^2 scores of the models that are compared.

2.10 Generalization Error and Overfitting

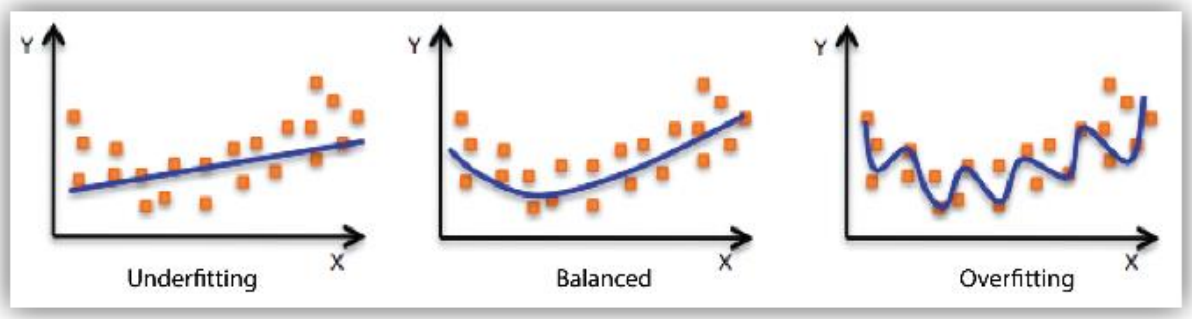

Figure 2.3: Example of underfitting, a balanced, and overfitted regressions. From Amazon Machine Learning Developer Guide (AML, 2021)

Overfitting occurs when the model performs well on training data but not on new data presented to it. A model will be overfitted if it is able to memorize the data and unable to generalize the data. Underfitting is when the model performs poorly on the training data, meaning it is not able to gain a general understanding. This will occur if not enough training data is available or poor parameters are chosen (AML, 2021). An example of over- and underfitting can be seen in Figure 2.3.

Generalization error is a measure for how well a model has been able to gain a generalized understanding. It is defined as the difference between the empirical loss of the training set and the expected loss of a test set. This is in practice measured by the error difference in the training and test datasets. When a model does not memorize the training data, but rather is able to learn the underlying patterns, it is said to have good generalization.

The generalization error cannot be calculated since the expected loss probability distribution is not known to the learning algorithm. It can however be approximated using the test loss. The equation for the approximated generalization error (GE) using the loss function L is:

$$GE(f, s_N, t_{N_{test}}) \triangleq \left| L_{emp}\left(y_i, f(x_i)\right) - L_{test}\left(y_i, t_{N_{test}}\right) \right|, \qquad \text{Equation 11}$$

where L_{emp} is the empirical test loss function defined as:

$$L_{emp}(f, s_N) \triangleq \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f(x_i)\right), \qquad \{(x_i, y_i)\}_{i=1}^{N} \in s_N, \qquad \text{Equation 12}$$

and L_{test}, the empirical test loss is given by

$$L_{test}(f, t_{N_{test}}) \triangleq \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} L\left(y_i, f(x_i)\right), \quad \{(x_i, y_i)\}_{i=1}^{N_{test}} \in t_{N_{test}}. \qquad \text{Equation 13}$$

In these equations x_i are the explanatory variables and y_i corresponding the target value in the training set (s_N) and the test set ($t_{N_{test}}$). These sets contain N and N_{test} data points respectively. $f(x_i)$ is the model prediction for the i-th measurement. (Jakubovitz *et al.* 2019)

The loss function L in a regression problem can for example be the RMSE or the MAE (Qi *et al.,* 2020). In the results section the MAE has been used as the loss function to quantify the generalization error. It can be seen as a metric for how much the model is overfitting. Keeping some data out of the training dataset is called bootstrapping, and is a feature included in several models. In this paper, block bootstrapping has been used for picking the test set. The blocks are all the measurements from the individual stations, and the test set is created from a portion of these blocks.

3. Experimental protocol

This section covers the data sources used for input variables in section 3.1. In section 3.2 the way these variables were constructed is outlined and section 3.3 cover the parameter selection for the models.

3.1 Study area

The study is conducted using data collected from Canada and the United States, with two separate validation datasets. The first of the validation datasets consists of CHSS and SNOTEL stations that are also used in the training of the models. The second consists of stations outside the area where the model was trained. This validation dataset consists of the SNOTEL stations in Alaska. The reason for validating the model twice is to get a better impression of the generality of the model findings, and to explore how well performing the models will be at predicting SWE in areas lacking SWE measurements.

While an extensive record of historical *in-situ* snow depth and SWE measurements exists, other meteorological variables have been estimated. The ERA5-Land dataset is a reanalyzed dataset using historical weather observations from the entire globe. All these observations are combined with modeled weather systems to create an hour-by-hour dataset of historical meteorological variables. (Sabater *et al.,* 2021)

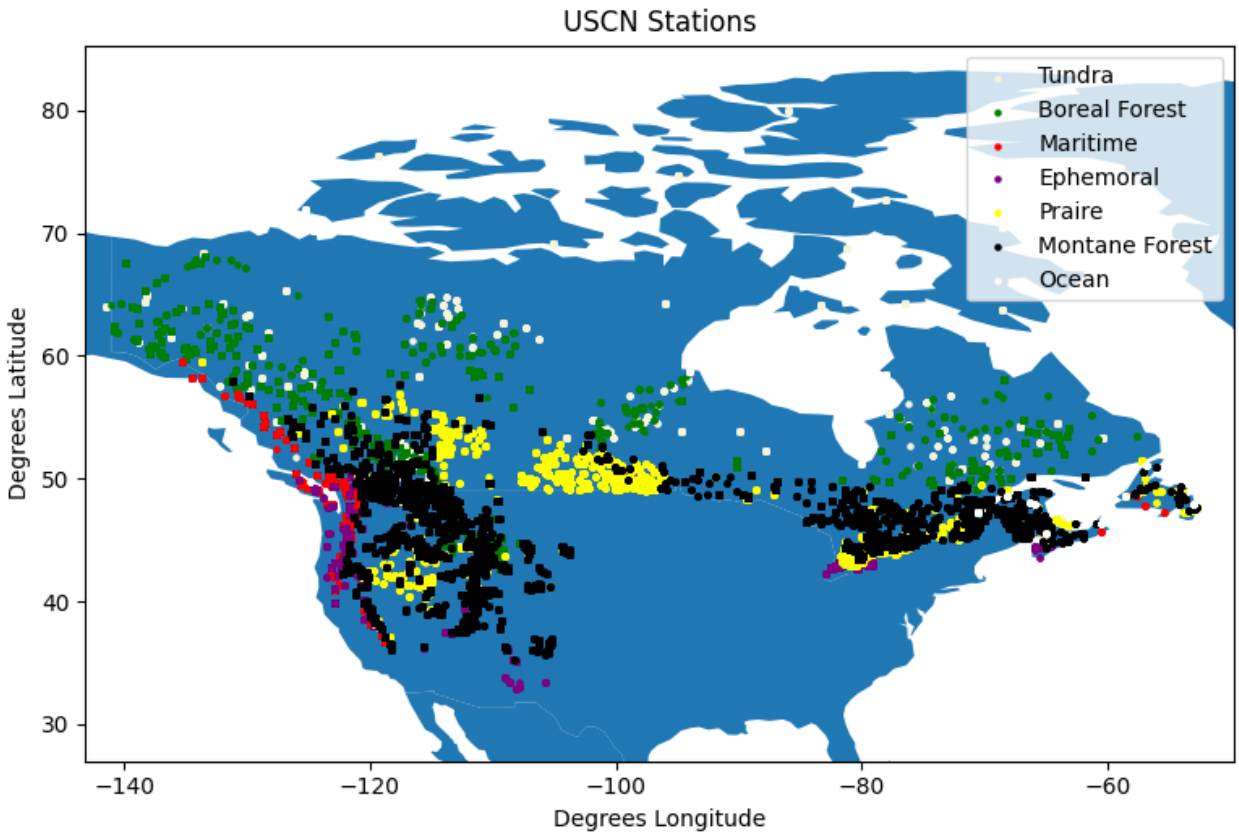

Figure 3.1: Map over Canada and the contiguous United States showing the location and snow class for stations in the USCN dataset.

3.1.1 Contiguous United States and Canada

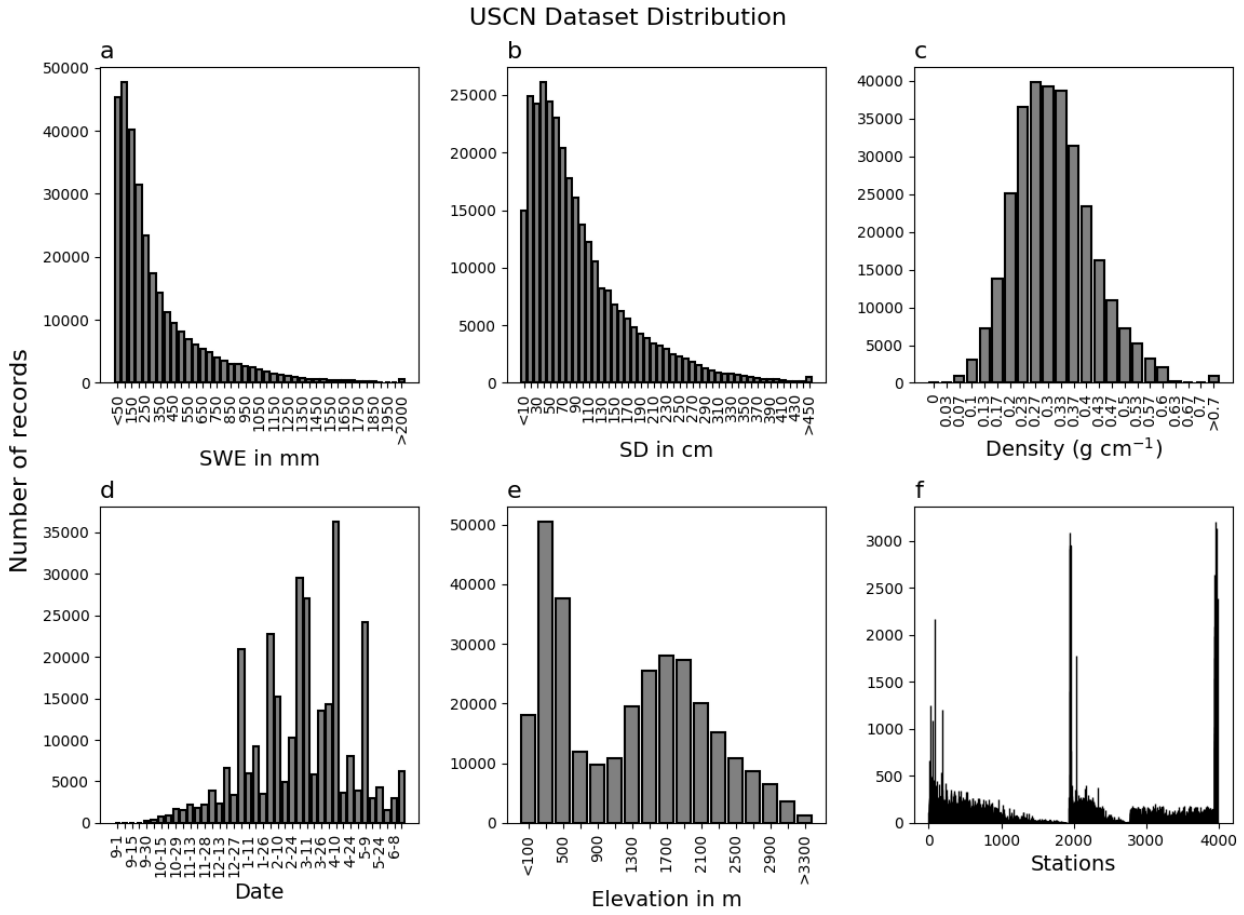

Figure 3.2: The USCN distribution of measured SWE depths at 50mm intervals. b) USCN distribution of measured snow depth at 10cm intervals. c) Distribution of USCN density calculated using Equation 1 shown with 0.03g cm^{-3} intervals. d) Temporal distribution of USCN measurement dates at 7-day intervals. e) Elevation of USCN stations rounded down at 200m intervals. f) Number of measurements obtained from each USCN station.

The Canadian Historical Snow Survey (CHSS) dataset is used for the Canadian measurements. 231692 measurements from 2821 stations come from the CHSS dataset. The US Snow Telemetry (SNOTEL) dataset is not used in its full form due to changes in measurement frequency. One measurement was collected for each 7-day period for stations with frequent reporting, although most stations only reported once per month. 1185 SNOTEL stations from the contiguous United States were used with a total of 78350 measurements collected. The combination of the Canadian and contiguous United States measurements is hereafter referred to as the USCN dataset. The locations of the USCN stations, as well as what snow class they belong to, can be seen in Figure 3.1.

In all, 310042 SWE measurements have been collected from 4006 meteorological stations covering the period from 01.01.1980 to 01.01.2019. Of these, 10840 measurements were removed due to false and dubious measurements, leaving a total of 299202 measurements. The various distributions of the dataset can be seen in Figure 3.2.

Table 3.1: Snow class distribution of the USCN dataset

Snow class	Measurements	Stations	Percentage (%)
Maritime	27476	257	9.18
Montane Forest	147066	1895	49.15
Ephemeral	7660	176	2.56
Prairie	42323	778	14.15
Boreal Forest	53451	653	17.86
Tundra	19593	215	6.55

As the study area covers large parts of the North American continent, many different climates are included. Table 3.1 shows the snow class distribution, the number of measurements and number of stations from each snow class. The snow class distribution seen in Table 3.1 shows that Montane Forest is the largest with 49.15% of the measurements and the Ephemeral the smallest with only 2.56%.

The snow depth distribution shown in Figure 3.2a has a mean depth of 90.3 cm and a median depth of 69.0 cm. The snow depths in the USCN dataset have a right-skewed gamma distribution with a skewness of 1.61. The SWE measurements seen in Figure 3.2b has a mean depth of 302.1 mm and a median depth of 183.0 mm. The SWE gamma distribution is also right-skewed with a skewness of 2.15.

The density distribution seen in Figure 3.2c is normally distributed with values ranging from 0.05 to 0.70 g/cm^3. The skewness of the density gamma distribution is 0.51. The mean density for the USCN dataset is 0.299 g/cm^3 while the median density is 0.289 g/cm^3.

The elevation distribution of the stations can be seen in Figure 3.2e. The median elevation is 1288 m and the mean elevation for all measurements is 1194 m.

Several spikes can be seen in Figure 3.2d, these are the 7-day periods containing the end of the month when many SNOTEL measurements have been gathered.

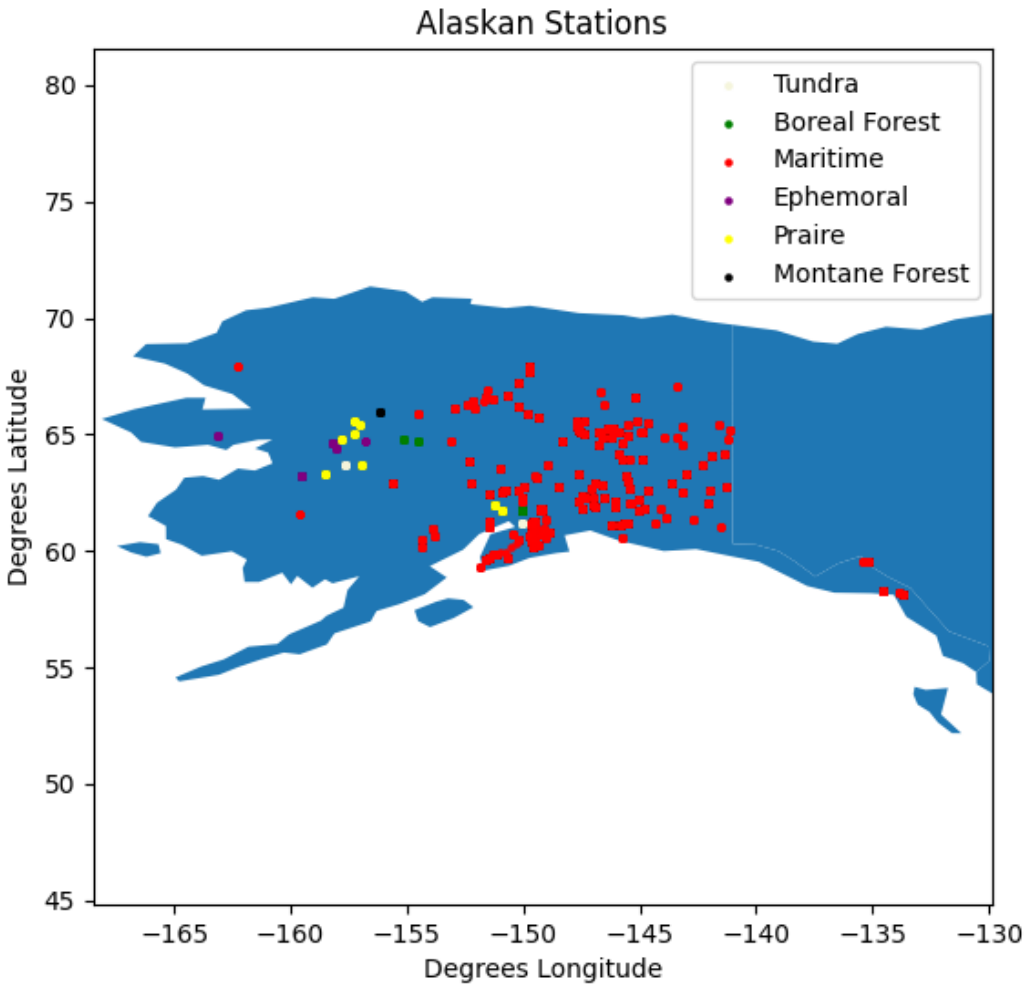

Figure 3.3: Map of Alaskan SNOTEL stations showing their location and snow class.

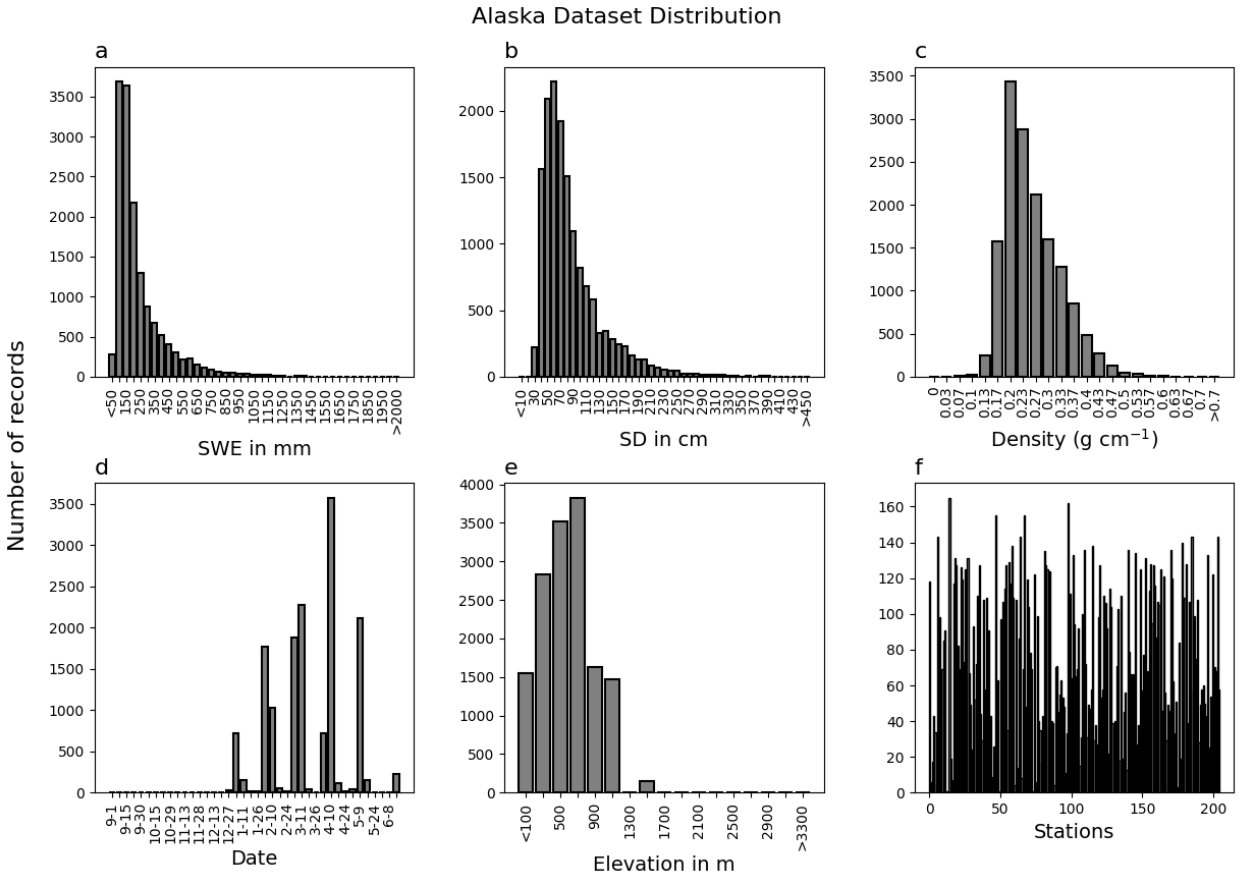

Figure 3.4: a) The Alaskan distribution of measured SWE depths at 50mm intervals. b) Alaskan distribution of measured snow depth at 10cm intervals. c) Distribution of density in Alaskan measurements calculated using Equation 1 shown with 0.03g cm^{-3} intervals. d) Temporal distribution of Alaskan measurement dates at 7-day intervals. e) Elevation of Alaskan stations rounded down at 200m intervals. f) Number of measurements obtained from each Alaskan station.

3.1.2 Alaska Dataset

The Alaskan dataset is a part of the SNOTEL dataset consisting of 14985 measurements from 211 stations. It was set aside so that the models could be cross validated using a block bootstrap approach. The Alaskan dataset was chosen for its uniqueness as it is measured with the SNOTEL methodology (USDA, 2022) while being geographically closer to Canadian stations. The different distributions of the Alaskan dataset differ from that of the contiguous United States and Canada, which can be seen in Figure 3.3. The distribution between snow classes for the Alaskan dataset can be seen in Table 3.2. A big majority of the measurements are situated in the Maritime snow class with 93.13% of the measurements. The Montane Forest and Tundra classes are hardly represented at all with less than 1% of the measurements. All the Alaskan station locations, as well as their snow class, can be seen in Figure 3.4.

Table 3.2: Snow class distribution of the USCN dataset

Snow class	Measurements	Stations	Percentage (%)
Maritime	13955	185	93.13
Montane Forest	31	1	0.21
Ephemeral	160	5	1.07
Prairie	519	9	3.46
Boreal Forest	201	3	1.34
Tundra	118	2	0.79

The Alaskan snow depth measurements have a mean depth of 80.5 cm and a median value of 66.0 cm. The Alaskan snow depths have a right skewness of 1.93. The SWE measurements have a mean depth of 212.7 mm and median depth of 147.0 mm. The SWE measurements are right skewed with a skewness of 2.52. This makes the mean snow depth 9.8 cm shallower than the USCN dataset while the mean SWE is 89.5 mm lower. The Alaskan snow depth and water equivalents are more right skewed than the USCN dataset.

The density distribution of the Alaskan dataset differs greatly from that of the USCN. The Alaskan density distribution is normally distributed with a skewness of 0.89. The Alaskan densities have mean and median values of 0.240g/cm^3 and 0.224g/cm^3 respectively. The densities in the Alaskan dataset are lower than the USCN with the mean density being 0.059g/cm^3 lower.

The median elevation of the Alaskan stations is 472 m and the mean elevation for all measurements is 487 m. This makes the mean elevation for the USCN dataset 707 m higher than the Alaskan.

The Alaskan dataset was divided into test and validation subsets. The two subsets were created splitting along the stations, with 106 stations and 7504 measurements in the validation dataset and 105 stations and 7481 measurements in the test dataset.

3.1.3 Removing outliers and false measurements

Several stations measurements were of suspicious quality, with an example station shown in Figure 3.5. These outliers were detected by calculating the snow density and manually reviewing outlier cases. The criteria being cases the densities exceeded 0.7 or where density was below 0.05 and depth over 1m. All measurements meeting these criteria were removed as these errors likely stem from errors in the measuring equipment. Several possible causes for these outliers are mentioned in Section 1.2.

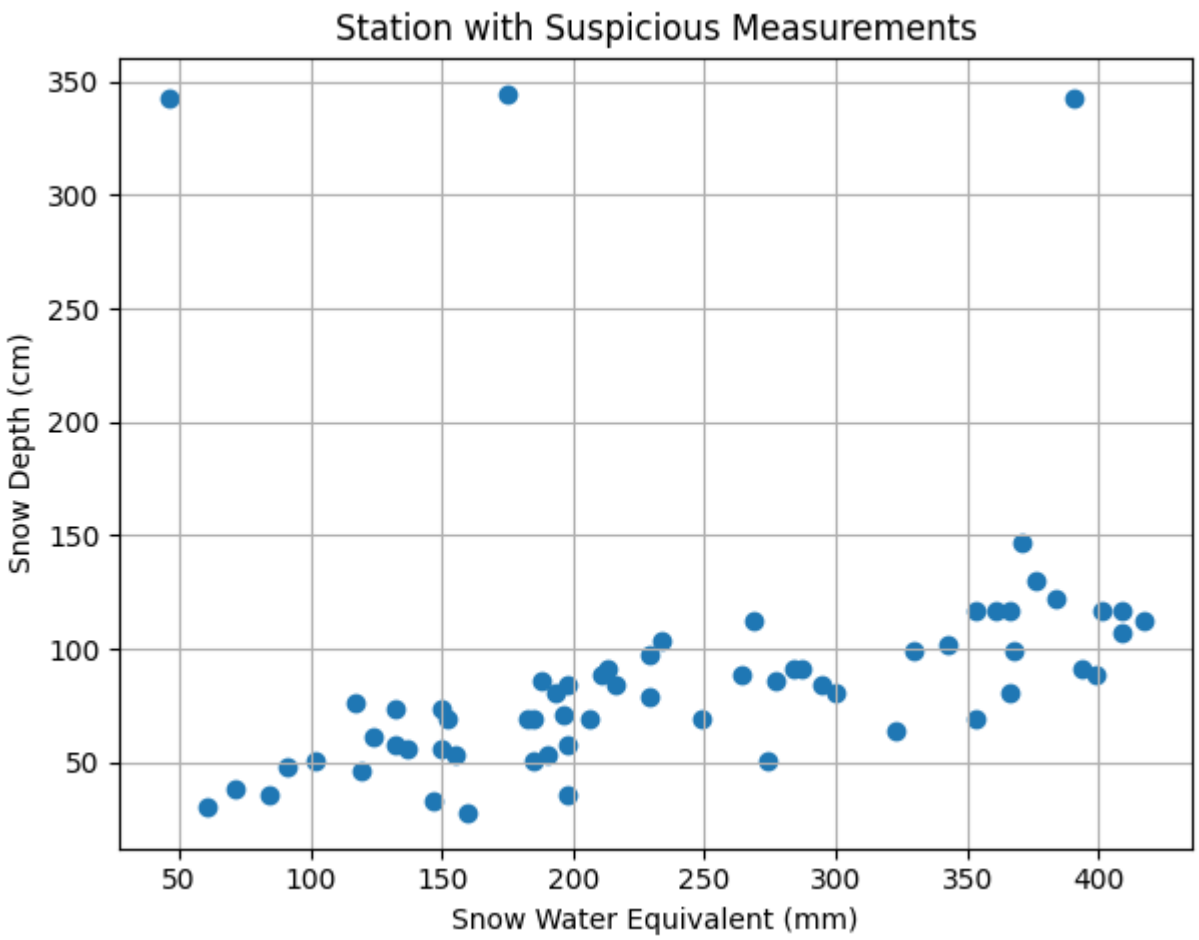

Figure 3.5: Example of CHSS station 18E009S which contains three dubious measurements seen lying far above the rest of the measurements in terms of snow depth.

3.2 Explanatory variables

From the ERA5-Land temperature and precipitation data several further explanatory variables can be constructed. For this study, two variable datasets were collected from the ERA5-Land dataset, these being hourly temperature and hourly precipitation. The datasets were processed in order to get the daily maximum and minimum temperature, as well as the daily precipitation for each grid cell where snow measurements existed.

The variables used are based on those proposed in Odry *et al.* (2020). Ntokas *et al.* (2021) used the same parameters but added snow density from ERA5-Land, which also proved to be their lowest scoring variable in terms of predictive power. Since this would require further ERA5-Land datasets to be acquired, the inclusion of this parameter was decided against. The snow classes in this paper are used as an explanatory variable, while earlier papers have trained separate models for each class. The explanatory variables used are:

- Snow depth from *in-situ* observations.
- Day of year, days since September 1st.

- Days without snow, the accumulated sum of days without meaningful snowfall since September 1st.

$$DWS = \sum_{i=-122}^{DOY} 1 \rightarrow (Pr > 3mm \; and \; T_{max} < 0)$$

- Number frost-defrost cycles. Using daily min and max temperature, it is calculated as the number of times the temperature fluctuates past the freezing/thawing threshold. These directional thresholds are set to −1C° and +1C° for the min and max temperature respectively.
- Number of days with positive degrees, the sum of the number of days between September 1st and the date of measurement where the max temperature exceeded zero degrees.
- Snow-cover aging index, the mean number of days since the last occurrence of solid precipitation weighted by the total solid precipitation on the day it fell. The weights range from 0 on September 1st and 1 on the day of the measurement (DOY).

$$SCAI = \frac{\sum_{i=1}^{DOY+122} \frac{i}{DOY+122} Pr_i}{\sum_{i=-122}^{DOY} Pr_i}$$

Where Pr_i is solid precipitation on the i-th day.

- Number of layers, estimated from the intensity of solid precipitation. A new layer is defined as a gap three days or longer where the daily calculated solid precipitation does not exceed 3 mm.
- Accumulated solid precipitation since the beginning of the season.
- Accumulated solid precipitation in the last 10 days.
- Total precipitation in the last 10 days.
- Average temperature in the last 6 days.
- Snow class according to Sturm *et al.* (2021).

The temperature has been corrected for using Equation 6 following Ntokas *et al.* (2021).

3.3 Tested Hyper Parameters

Selecting the right hyper-parameters (parameters) is an important step when constructing a model. A poor parameter selection can lead to a model overfitting or underfitting depending on the parameters used (AML, 2021). The CART models have many parameters to tune how the model trains on the training data. All the parameters were tested using the *ceteris paribus* principle like in Ntokas *et al.* (2021) where the tested parameter is changed while all other parameters are kept the same. The XGBoost (XGB) model and the Random Forest (RF) model do not have identical parameters as they are constructed differently. The hyper-parameter names used in the scikit-learn packages are also different, even for the parameters that are similar (Buitinck *et al.* 2013). A common nomenclature has been used in this paper when referring to similar parameters.

For the XGB model, the parameters tested were the max depth of the individual CARTs (layers), the number of estimators (estimators), the learning rate, and the number of explanatory variables considered

when training each tree (max features). It was decided to refer to the max depth of the CARTs as layers to avoid confusion with snow depth or SWE depth.

For the RF model the parameters tested were the max depth (layers), number of trees (estimators) and max features. These are the parameters with common nomenclature to the XGB model. The minimum number of samples required to make a split and the minimum number of samples in the leaf node were additionally tested for the RF model. The max features parameter value "None" will mean that the number of features considered equals the number of variables. This means that all explanatory variables are considered when constructing each CART.

Table 3.3: Best MLP Parameters from Ntokas *et al.* (2021)

Parameter	Used
Activation function	tanh
Optimization algorithm	AdaDelta
Parameter initialization	$U(-2, 2)^3$
Shuffling data before epoch	Yes
Input variables	All
Batch size	100
Ensemble members	20
Number of ensembles	6

The tested parameters for the MLP model are the number of hidden layers and number of epochs. The other parameters used are those found to give the best results in Ntokas *et al*. (2021) seen in Table 3.3.

In this paper the term complex is used for a model where many layers and estimators are used during the training phase while simple models are used when few are used.

4. Results

The results section consists of two similar approaches to obtain model SWE predictions. In section 4.1 the test dataset stations are represented in the training dataset, while section 4.2 follows a block bootstrap approach where the stations are separated. In section 4.2 the ubiquitousness of the findings is further explored by repeating the method for different areas and dataset divisions. In section 4.3 the feature importance of the explanatory variables is presented, and section 4.4 explores the computational costs of the models.

4.1 USCN Validation

For finding the best parameters for the MLP, RF, and XGB models, the USCN dataset was divided into three separate datasets. 50% of the dataset is used for training the model, 25% for validating the model and finding the best parameters, and 25% for testing the final model. For the regression models the USCN dataset was split into two parts where 75% was used for training and 25% for testing.

4.1.1 Training

All the snow classes were used in the training of the CART models, leaving the snow class as a categorical feature for each data point.

4.1.1.1 XGBoost

Table 4.1: Parameter ranges used when training the XGB model

Parameter	Best	Tested
Number of estimators	1500	[10, 1500]
Layers	9	[2, 11]
Learning rate	0.1	[0.001, 0.5]
Max features	1	[0.5, 1]

The XGB model was trained with the parameters from section 3.3 being optimized. The parameters were tested one by one against a reference parameter combination and validated after each iteration. The best values for all parameters can be seen in Table 4.1. The learning rate and max features were found using this technique. These parameters were found using 7 layers and 500 estimators as reference parameters. After other appropriate parameters were found, the optimal numbers of estimators and layers were determined through a grid search. Using 161 different numbers of estimators in the range [1,1500] and layers in the range [2,11], a total of 1510 combinations were tested. The results of tuning layers and number of estimators can be seen in Figure 4.1.

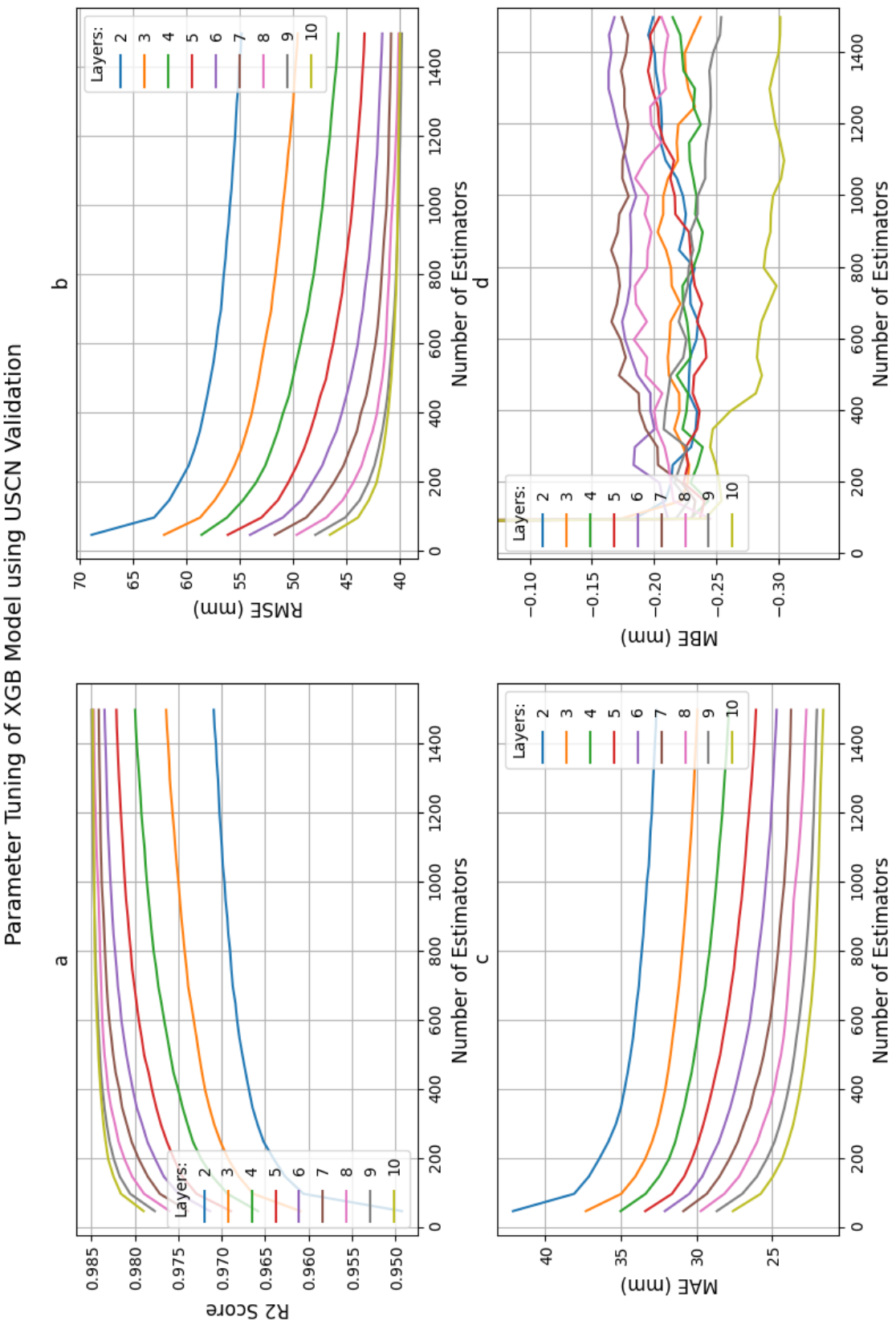

Figure 4.1: Performance of different XGB parameter combinations. Accuracy scores are plotted for each depth as a function of number of estimators. The metrics are improving with more layers and as the number of estimators increase. The metrics plotted are A) R^2 score, b) the RMSE in mm, c) MAE in mm, and d) MBE in mm.

The XGB performs better when adding complexity. Figure 4.1 shows that adding more layers and estimators improves the performance for (a) R^2, (b) RMSE and (c) MAE. (d) The MBE is not affected in the same way with increasing complexity, with most iterations having an MBE in the range -0.15 mm to -0.3 mm. The MBE difference between layers widens as more estimators are added, with the better performing models having a more negative bias. At 8 layers and 600 estimators the R^2 score reaches 0.9845 and the addition of more layers and estimators have less of an impact. The addition of more layers is the most computationally costly parameter of the model, with execution times increasing exponentially. As the depth and the number of estimators are increased, the model appear to converge on an R^2 score just shy of 0.985. The best layer and number of estimators parameters were 9 and 1500 respectively.

4.1.1.2 RF model

Table 4.2: Parameter ranges used when training the RF model

Parameter	Best	Tested
Number of estimators	300	[10, 300]
Layers	30	[4, 30]
Max features	None (all features included)	Sqrt, log2, None
Minimum number of samples (split)	3	[2, 10]
Minimum number of samples (leaf)	1	[1, 5]

The parameters for the RF model were found in the same manner as for the XGB model. The parameters tested can be seen in Table 4.2 along with the best parameters. The depth and number of estimators was then found by grid search. 7 different numbers of layers in the range [4, 30] were each trained in combination with 41 different numbers of estimators in the range [1, 300]. The resulting R^2 score, RMSE, MAE, and MBE progression can be seen in Figure 4.2. The parameters resulting in the best performance were 30 layers and 300 estimators.

The results follow the same pattern seen while training the XGB model with improving performance as complexity increases. In Figure 4.2 it can be seen that: for iterations with less than 12 layers, the model is not able to reach an R^2 score above 0.98 (a). For models constructed with 6 or more layers this barrier is reached with less than 400 estimators. The RMSE (b) and MAE (c) follow similar downward trends with increasing complexity. The MBE is not as stable between iterations but is found in the range [-0.15 mm, -0.25 mm] for more complex iterations (d).

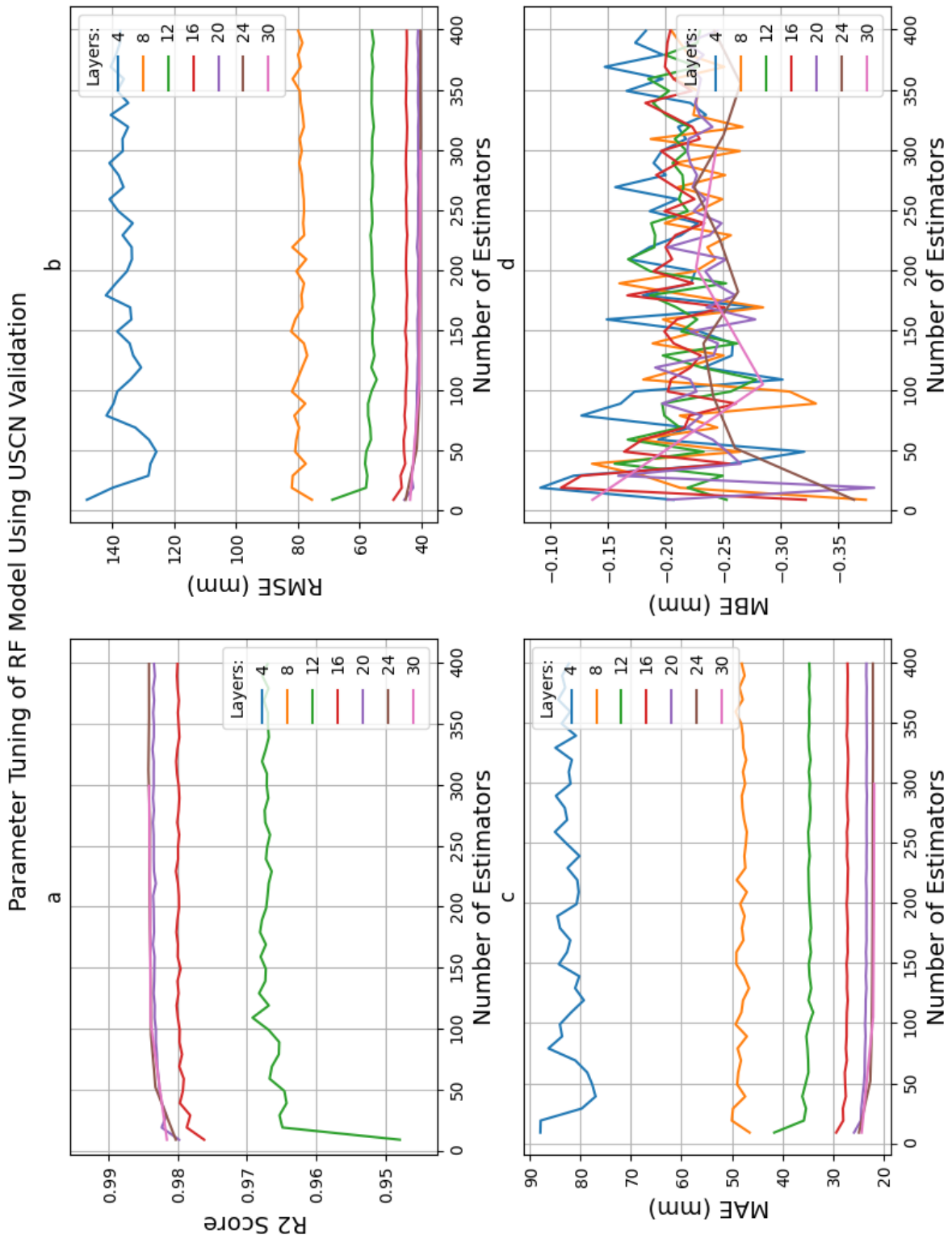

Figure 4.2: Performance of different RF parameter combinations. Accuracy scores are plotted for each depth as a function of number of estimators. The metrics are improving with more layers and as the number of estimators increase. The metrics plotted are A) R^2 score, b) the RMSE in mm, c) MAE in mm, and d) MBE in mm.

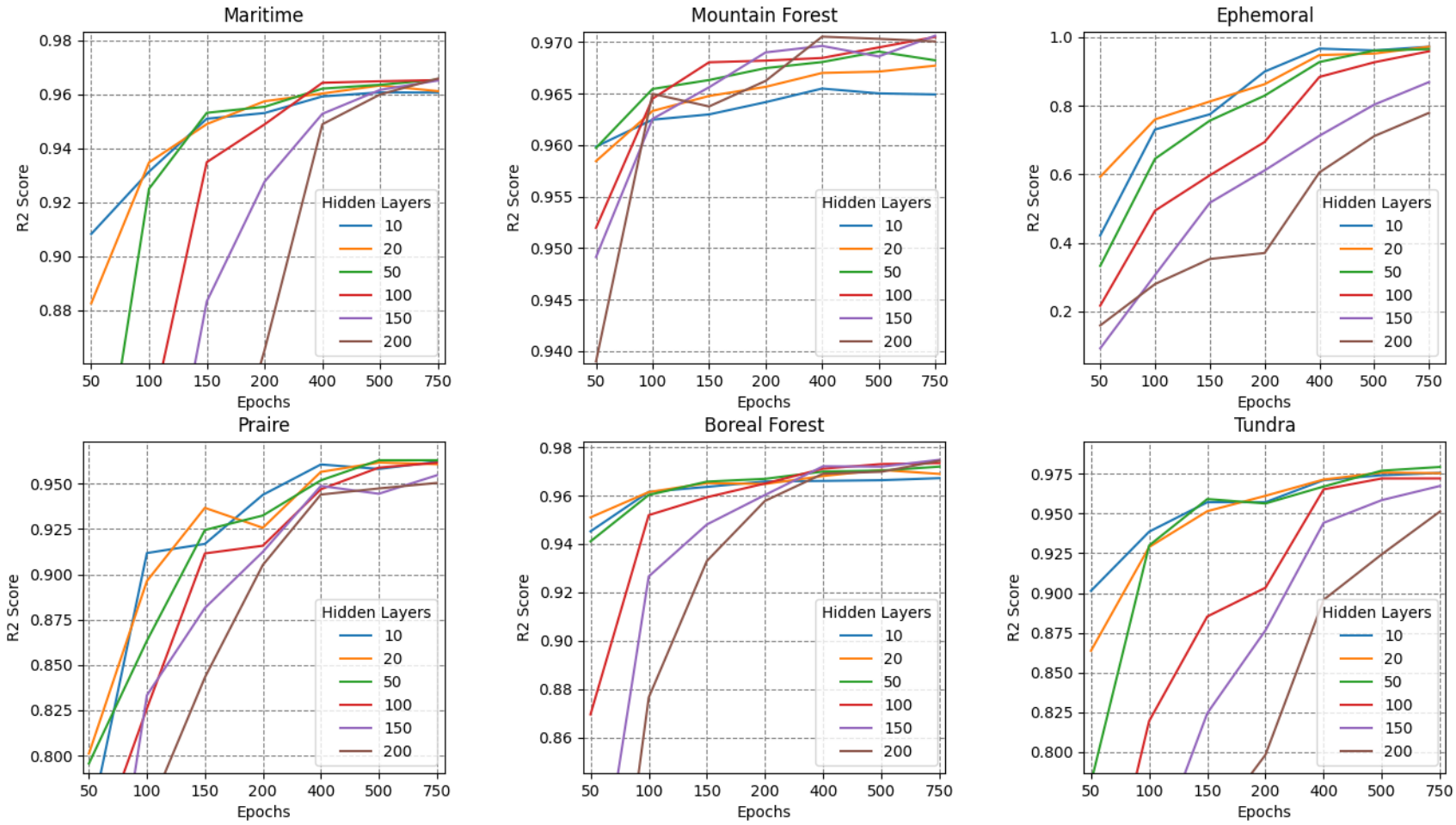

Figure 4.3: Testing of the MLP model parameters for the USCN dataset for the six different snow classes. Each plotted line represents a number of hidden layers and the R^2 score is a function of increasing number of epochs.

4.1.1.3 MLP

Table 4.3: Best Epoch and Hidden Layer parameters found for the MLP model as well as tested ranges

Snow class	Epoch	Hidden Layers
Maritime	400	100
Montane Forest	400	200
Ephemeral	400	10
Prairie	500	50
Boreal Forest	750	50
Tundra	750	150
Tested Range	[50, 750]	[10, 200]

The MLP model was trained with the parameters tuned being the number of epochs and the number of hidden layers. For the parameter tuning, only one ensemble member was used to save time. The tuning was done for each snow class individually, creating a grid to find the best parameters within the ranges shown in Table 4.3. For each snow class 36 combinations of epochs/layers were tested. Ntokas *et al.* (2021) tested for many parameters when constructing their model. The parameters that gave the best results in their study were used here and can be seen in Table 4.3. The best epoch and hidden layer parameters for each snow class can be seen in Table 4.3 and the R^2 scores from the different combinations used for training can be seen in Figure 4.3.

4.1.1.4 Regression models

The Sturm and Jonas models were implemented as described in Section 2.4 and 2.5 respectively. The Sklearn logistic regression package was used to find the optimal parameters. The regression models were trained using the snow density as the target which was then converted to SWE in order to run a comparative performance analysis. For the Jonas model, some elevation and month combinations had too few data points to make a proper regression fit. Therefore, a lower boundary was set and only month and elevation combinations with at least 10 measurements were considered. The months without enough data points were all summer months.

4.1.2 Performance

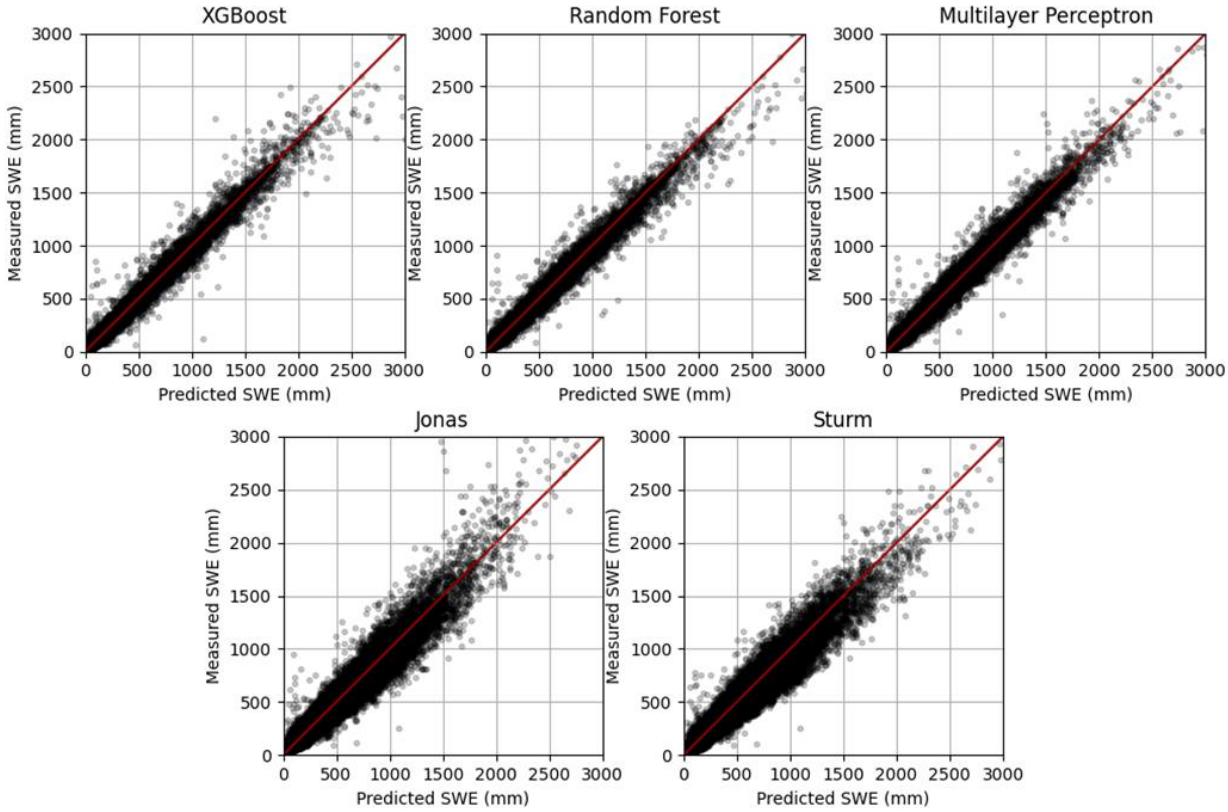

Figure 4.4: Scatter plot of measured values (mm) and predicted values (mm) when validating for the USCN dataset. The red line marks where x=y which is the theoretical best predictions. Limits for the plot were set to 3000mm meaning that not all measurements are included in these plots.

Table 4.4: Model performance when applied on the USCN test dataset

Model	MSE (mm)	RMSE (mm)	MAE (mm)	MBE (mm)	R^2
XGBoost	1614.9	40.19	21.40	-0.32	0.985
Random Forest	1623.7	40.30	22.04	-0.25	0.985
Multilayer Perceptron	2127.6	46.13	28.52	0.09	0.980
Jonas	5287.4	72.71	43.50	-0.11	0.950
Sturm	5365.9	73.25	45.68	1.99	0.945

Both XGB and the RF model outperformed previously suggested SWE models. In Figure 4.4 all model predictions for the test data set are plotted, alongside the theoretical best fit. Better model performance will produce a narrower cluster of modeled points around the theoretical best fit line shown in red. The corresponding R^2, RMSE, MAE and MBE scores are found in Table 4.4. Comparing the results shows the Sturm model to have the weakest performance with an R^2 score of 0.936, followed by the Jonas model with 1.5% better R^2 score of 0.951. The more complex MLP model, with an R^2 score of 0.980 had a 3.1% increase in precision compared to the Jonas models. The CART models compared with the MLP model show an improvement of 0.48% and 0.49% for the RF model and XGB model respectively.

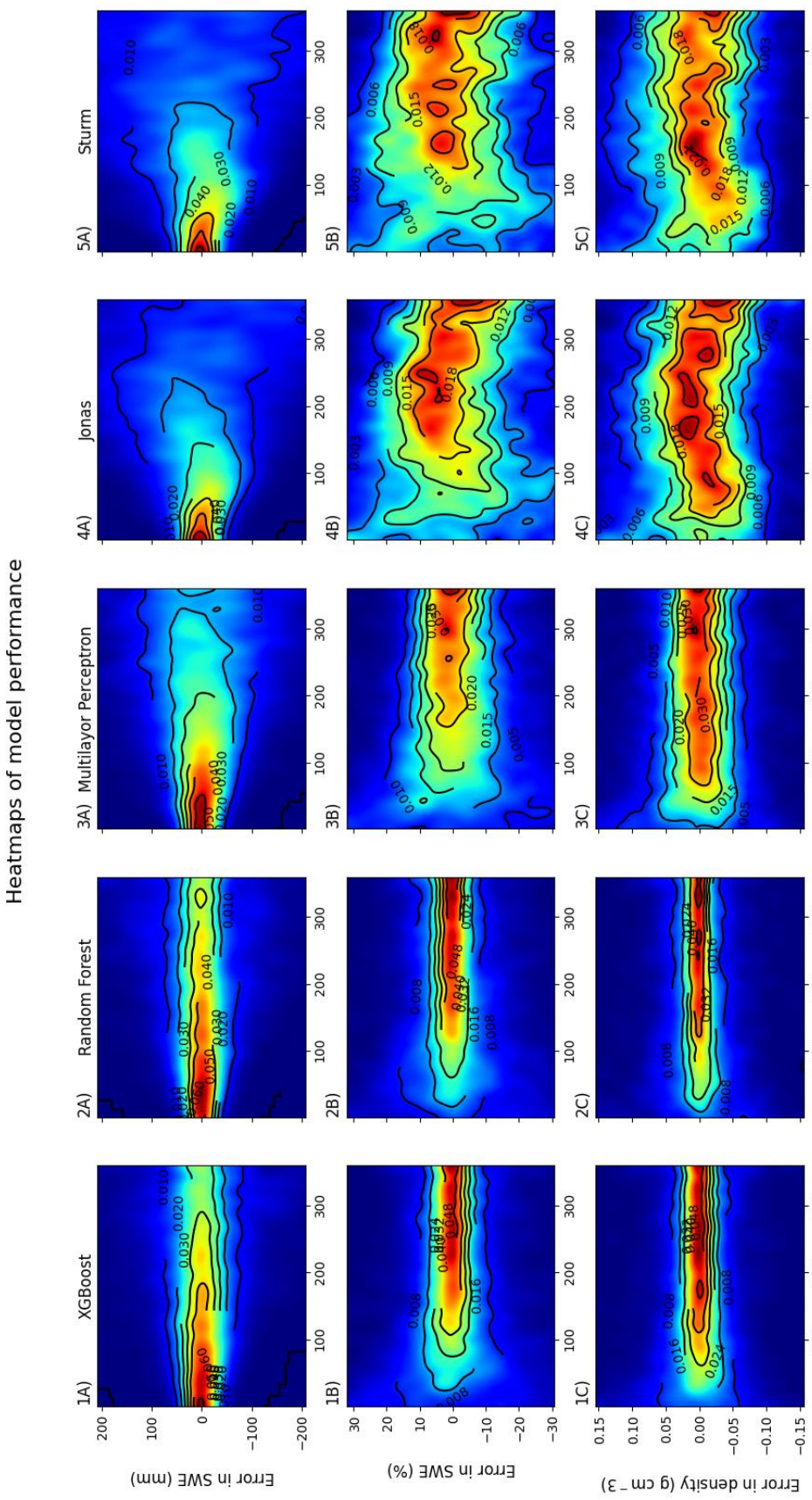

Figure 4.5: Heat maps of model performance for all models (1-5) on the USCN test dataset. Better performance will produce narrower bands of high error densities centered around zero. The measurements have been chosen so an equal number of measurements are represented for each 10cm depth interval. The errors shown are a) SWE error in mm, b) relative SWE error in percent, and c) density error in g cm^{-3}. The error densities are displayed as a function of snow depth in cm. The scale goes from deep blue where there is low density of prediction errors to dark red where there is highest error density.

The heatmaps displaying the probability of error as a function of snow depth in Figure 4.5 were created using 150 data points for each 10 cm snow depth interval in the range [0 cm, 350 cm]. A total of 5250 measurements were used, all obtained from the test dataset. The heat maps are smoothed out from a scatter plot, using a smoothing algorithm to convert the scatterplot to a probability density plot. Increasing model performance will be observed in these plots by narrower bands of error probability.

The CART models perform best for all snow depths as can be observed by the probability of error in Figure 4.5. Observing first the density error probability, the models with lower accuracy scores produce errors distributed over larger ranges. The regression models have their density errors +/- 0.1 g/cm^3 (4c, 5c), while the MLP model has errors of +/- 0.05 g/cm^3 (3c). The XGB and RF models have their errors in the +/- 0.025 g/cm^3 range (1c, 2c). The complex models all tend to have better predictions as the depth increases. While this is also true for the regression models it is much less pronounced in the density errors and is perhaps better illustrated in the relative error distribution (4b, 5b). The SWE error probability shows the opposite pattern to that of the density. All the models show greater SWE errors as snow depth increases as the larger values cause a greater absolute spread. The error graphs show the power of the CART models (1a, 1b) compared to the MLP model (3a) as the error probabilities throughout all depths are lower. This improvement is even better for snow depths > 200 cm. The relative SWE error shows the CART models have most of their predictions within a 10% range (1b, 2b), while the MLP is mostly within 15% (3b). The regression models are generally within 30% (4b, 5b).

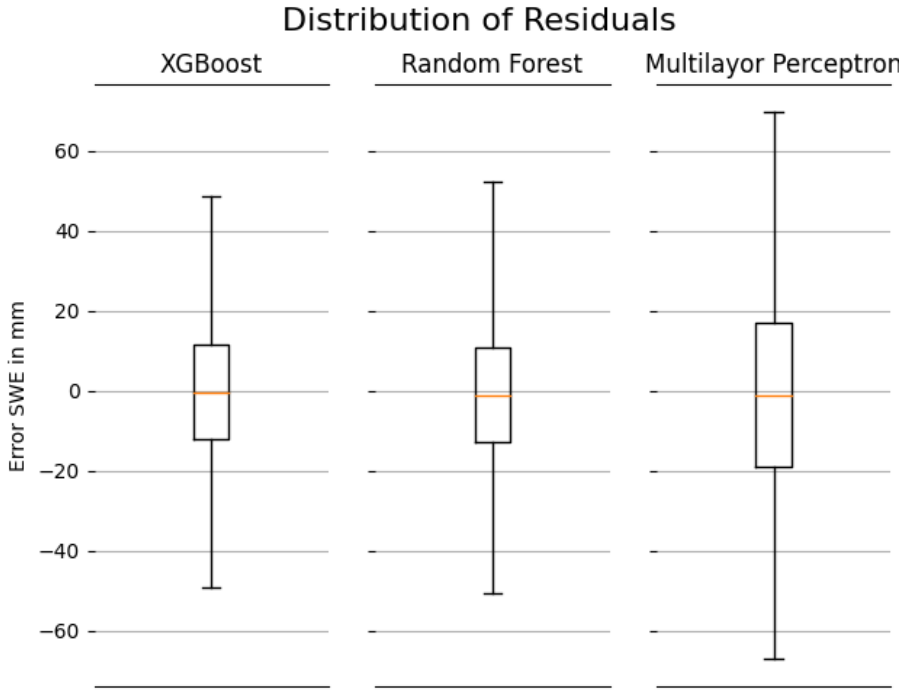

Figure 4.6: Box and whisker plot comparing the distribution of error residuals. The box and whiskers correspond to areas where percentiles of the errors are found within. The whiskers mark the 5th and 95th percentile, while the box shows where 50% of the residuals are found inside. The red line marks the mean of the residuals. The XGB, RF, and MLP models are compared as they had the greatest accuracy.

Figure 4.6 shows the distribution of the residual errors. The XGB model has 90% of its residuals within a range of [−49.1 mm, 48.7 mm] and 50% within [−12.0 mm, 11.4 mm] with a median of −0.3 mm. For the RF, 90% are within [−50.6 mm, 52.2 mm] and 50% within the [−13.0 mm, 10.6 mm] range with a median of −1.1 mm. The respective MLP residual ranges are [−64.3 mm, 68.1 mm] and [−16.7 mm, 16.9 mm] with a median of −0.8 mm. This shows a slight negative bias, which is smallest for the XGB model, followed by the MLP model and lastly the RF model.

Table 4.5: Model performance when applied on the Alaska dataset with generalization error calculated using MAE as the loss function in Equation 11.

Model	MSE (mm)	RMSE (mm)	MAE (mm)	MBE (mm)	R^2	GE(MAE)
XGBoost	4764.2	69.02	47.44	−35.46	0.870	26.04
Random Forest	3646.3	60.38	37.61	−24.98	0.908	15.57
Multilayer Perceptron	11185.4	105.76	85.89	−84.07	0.748	57.37
Jonas	8821.2	92.93	78.12	−83.51	0.761	34.62
Sturm	12079.2	109.91	95.69	−92.86	0.686	50.01

Table 4.5 shows that all the models have a big loss in performance when applied on the Alaskan dataset. This shows that there exists a considerable generalization error in all the complex models. This suggests that the models are overfitting and not learning the variable-SWE relationship. The RF model has the lowest generalization error of the models with a GE of -15.57 mm.

4.2 Alaska Validation

Due to the poor performance of the complex models when applied to the Alaskan dataset, it was decided to search for better parameters to get better performance and explore ways to minimize the generalization error. To search for better parameters, it was decided to split the Alaskan dataset into a validation and testing dataset, using a block bootstrap method where the measurements are divided along stations. This ensures that all measurements in the final test predictions will be from a site that the model has seen neither the training nor validation phase.

4.2.1 Training

Training the CART models were performed using the USCN training dataset and model parameters seen in Table 4.1 except for the number of layers and estimators. The Jonas and Sturm models was not retrained as their parameters are dependent exclusively on the training dataset.

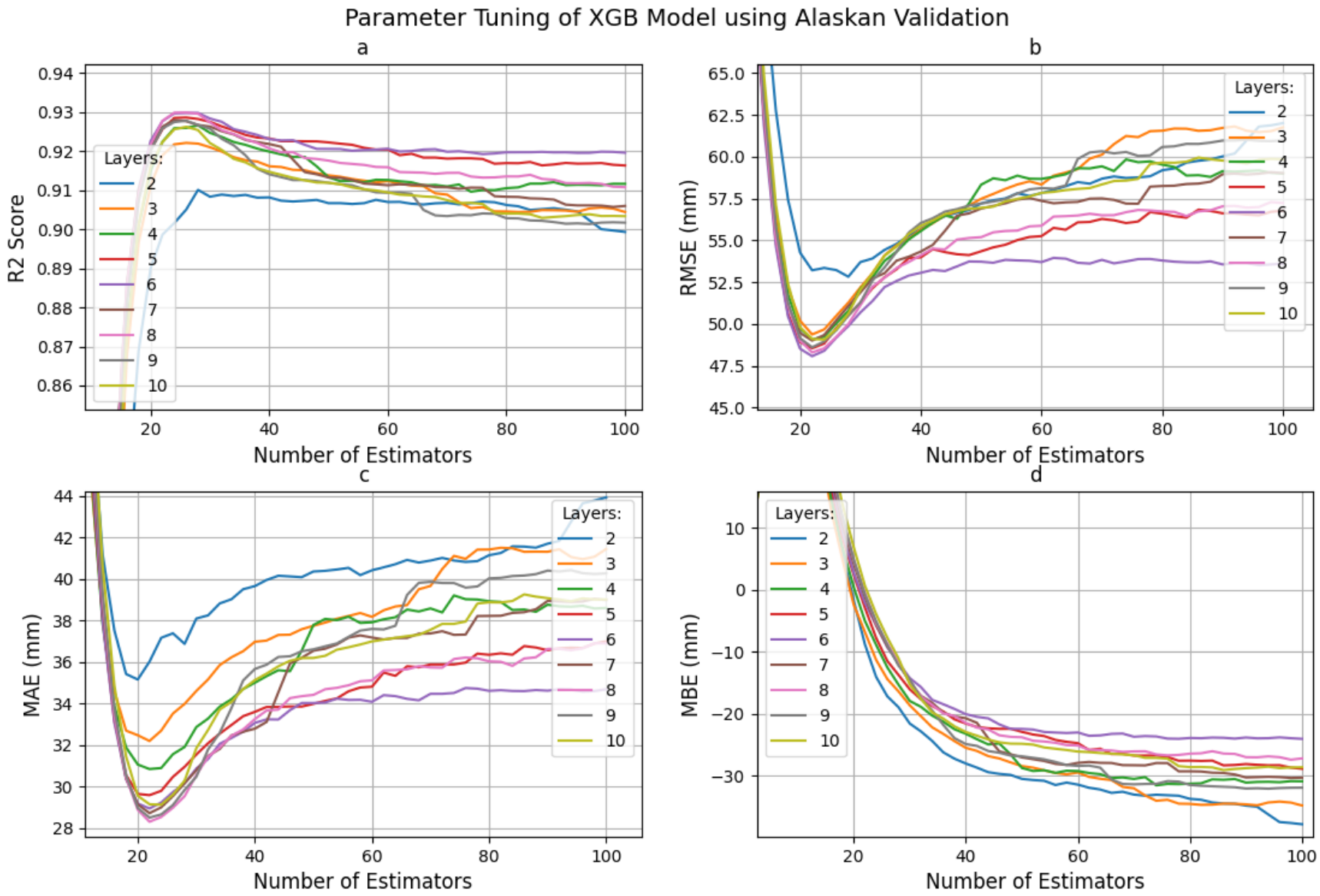

Figure 4.7: Performance of different XGB parameter combinations when validated for Alaskan stations. Accuracy scores are plotted for each depth as a function of number of estimators. The metrics plotted are a) R^2 score, b) the RMSE, c) MAE, and d) MBE.

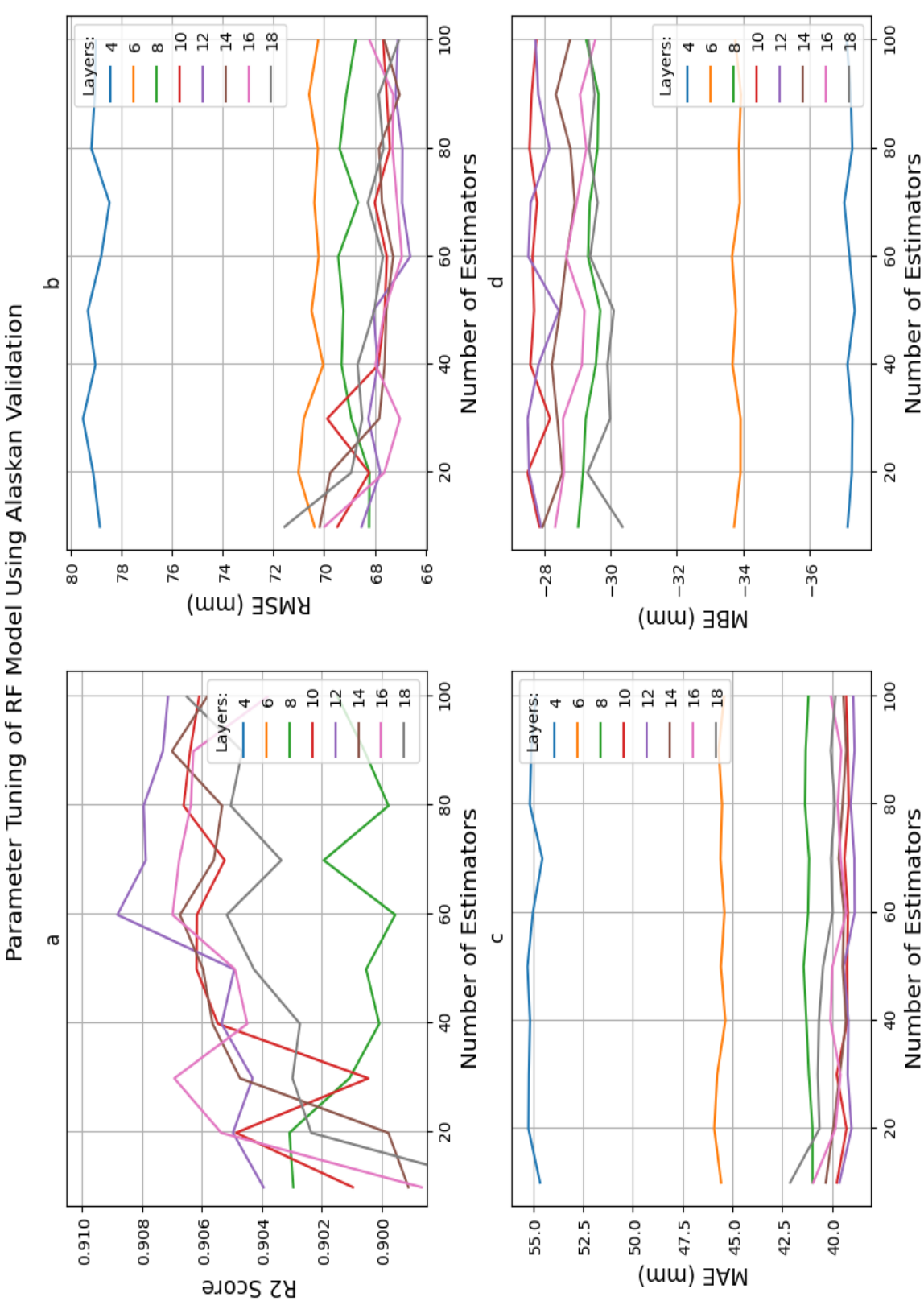

Figure 4.8: Performance of different RF parameter combinations when validating with Alaskan stations. Accuracy scores are plotted for each depth as a function of number of estimators. The metrics plotted are a) R^2 score, b) the RMSE, c) MAE, and d) MBE.

4.2.1.1 XGBoost

Optimizing the XGB model was done by running the model with several combinations of layers [2, 11] and estimators [10, 100]. The results of the XGB parameter search can be seen in Figure 4.7. The resulting R^2 score (a), MAE (b), and RMSE (c) of validating all the parameter combinations can be seen to follow each other in terms of wellness. The accuracy of the model is increasing with layers added until peaking at 7 or 8 before adding even more has a negative effect. The accuracy also increases with the number of estimators until the best results are reached at around 24. Subsequent addition of estimators only deteriorates model performance. The best performing combination of 7 layers and 24 trees gave an R^2 score of 0.93. The peak performance coincides with the MBE going from a positive bias to a negative bias, with the best validation parameters giving an MBE of -27.6 mm (d).

4.2.1.2 Random Forest

The RF model optimization was undertaken in the same way as the XGM model. Accuracy scores were produced by running the model for combinations of layers [4,20] and estimators [4, 100] seen in Figure 4.8. The RF model improves with more layers obtaining the best results with 12 layers before further addition shows deteriorating performance (a). The impact of adding estimators shows a general positive trend where more estimators give better performance. The best result of the validation is 12 layers and 60 estimators. the RMSE (b) shows the same trend as the R^2 score, while the MAE does not seem to be significantly impacted by the addition of estimators (c). The MBE plot reveals that all iterations had an MBE less than −27 mm (d). Figure 4.8 reveals that the number of layers is the most impactful parameter of the two when training the RF mode. The best iteration of 12 layers and 60 estimators had an R^2 score > 0.908 and the MBE closest to 0.

4.2.1.3 Multilayer Perceptron

When training the MLP model, difficulties arise around the distribution of snow classes in the Alaskan dataset described in section 3.1.2. Since all the classes except the Maritime contain less than 10 stations, a single MLP was constructed with snow class as an explanatory variable instead. Ntokas *et al.* (2021) tested a similar approach in their paper which proved to perform well, although not as well as the snow class specific ensemble. The MLP was then validated to find the best hidden layer and epoch parameters. The validation was performed on a single ensemble member with hidden layer [2, 200] and epochs [10, 100] combinations. The result of training the single MLP with these parameter combinations can be seen in Figure 4.9. Although it didn't produce the overall best result, 2 hidden layers were chosen when creating the final MLP with 20 ensemble members. This is due to it being the most consistent over several iterations of similar parameters.

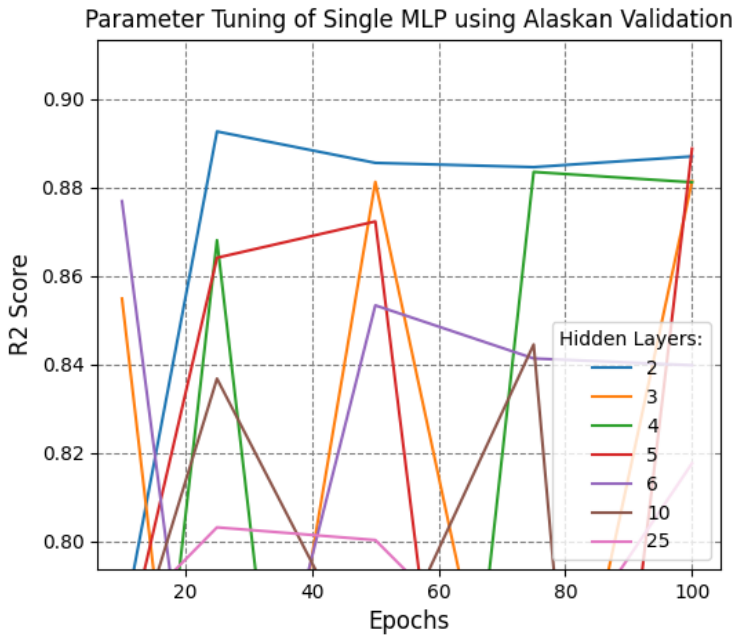

Figure 4.9: Testing of MLP model parameters for all snow classes combined conducted on 106 Alaskan stations. Due to the Alaskan dataset containing 93% Maritime measurements a single MLP was constructed taking snow class as a variable.

4.2.2 Performance

Table 4.6: Model performance when validated on the Alaskan validation set and applied on the Alaska test dataset. The generalization error is estimated with the MAE from this test data and the MAE from using the USCN test data.

Model	MSE (mm)	RMSE (mm)	MAE (mm)	MBE (mm)	R^2	GE(MAE)
XGBoost	1738.5	41.70	26.80	−3.57	0.934	12.03
Random Forest	2680.6	51.77	33.45	−22.10	0.914	2.29
Multilayer Perceptron	3225.4	56.79	39.12	−29.69	0.894	0.28
Jonas	7817.2	88.41	76.86	−72.85	0.731	33.36
Sturm	12079.3	109.91	95.59	−92.86	0.650	31.18

The measured values compared with the predicted values when the models are applied to the Alaskan test dataset can be seen in Figure 4.10. The R^2 score, RMSE, MAE, and MBE are found in Table 4.6. The XGB model has a 2.0% better performance than the RF model and a 4.0% better performance than the MLP model. The better performance of the XGB model can be seen in Figure 4.10 where its predictions are closer to the red line marking where the predicted values equal the measured ones. The Jonas model performs 8.1% better than the Sturm model. This is significantly better than when compared to the 0.05% difference in performance from table 4.4. The XGB model has a performance 20.3% better than that of the Jonas model. The generalization errors have been calculated for the XGB, RF, and MLP models by running the USCN test dataset with the same parameters for comparison. The use of early stopping for the XGB model causes a greater MBE when tested using these parameters found in 4.2.1 on the USCN test dataset.

The greater generalization error suggests that regional biases are being corrected for and thus a model bias exists.

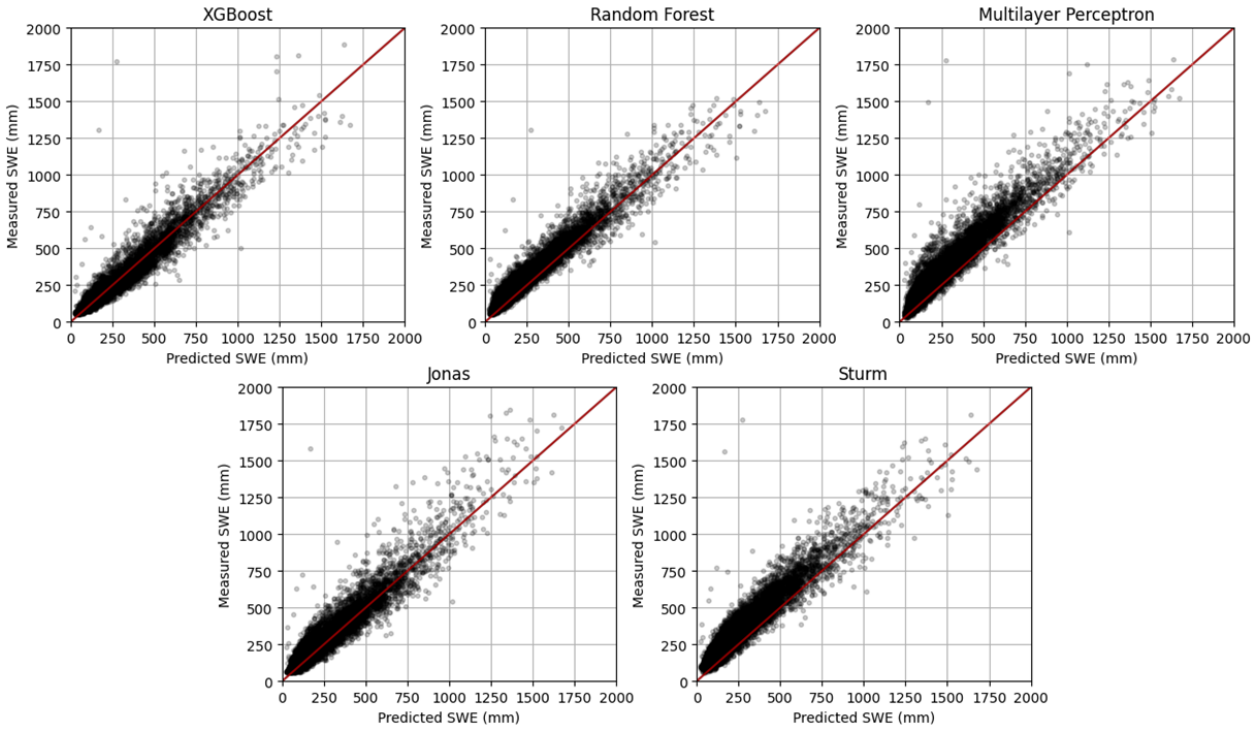

Figure 4.10: Scatter plot of measured values (mm) and predicted values (mm) when validating for the Alaskan dataset. The red line marks where x=y which is the theoretical best predictions. 2000 mm was selected as the limits, and thus not all measurements are included in the plot.

The heatmap in Figure 4.11 is created using the full Alaskan dataset with 100 measurements selected for each depth in the range [20 cm, 170 cm] containing measurements from both the testing and validation datasets to obtain enough measurements. 100 additional measurements in the range [170 cm, 500 cm] were added and normalized to be displayed between 170 cm and 180 cm. A recurring trend in all the models is underestimating the SWE.

The Sturm model shows the greatest tendency to underestimate. Figure 4.11-5 shows that as snow depth increases the area of high SWE error distribution goes from [−20 mm, 0 mm] at low snow depth to [−100 mm, 0 mm] at snow depth > 100 cm (a). The relative SWE error for the Sturm model is imprecise but improving as snow depth increases. At a depth of 100 cm a higher distribution of errors is seen at [−40%, −20%] moving towards [−30%, 0%] for snow depth > 170 cm (b). The density error for low snow depth has the highest probability at −0.1 g/cm^3. This decreases to −0.06 g/cm^3 as snow depth increases and the model gets more precise (c).

The Jonas model seen in Figure 4.11-4 shows much narrower error distributions than the Sturm model. The SWE error probability is centered around −20 cm at the lowest and spreads out as snow depth increases, being centered around −30 mm (a). The relative SWE error probability shows a similar trend as the Sturm model with imprecise predictions for snow depth < 100 cm (b). The density error has a high

probability for low snow depth at −0.06 g/cm^3 decreasing with snow depth and being centered around −0.01 g/cm^3 at snow depth > 100 cm (c).

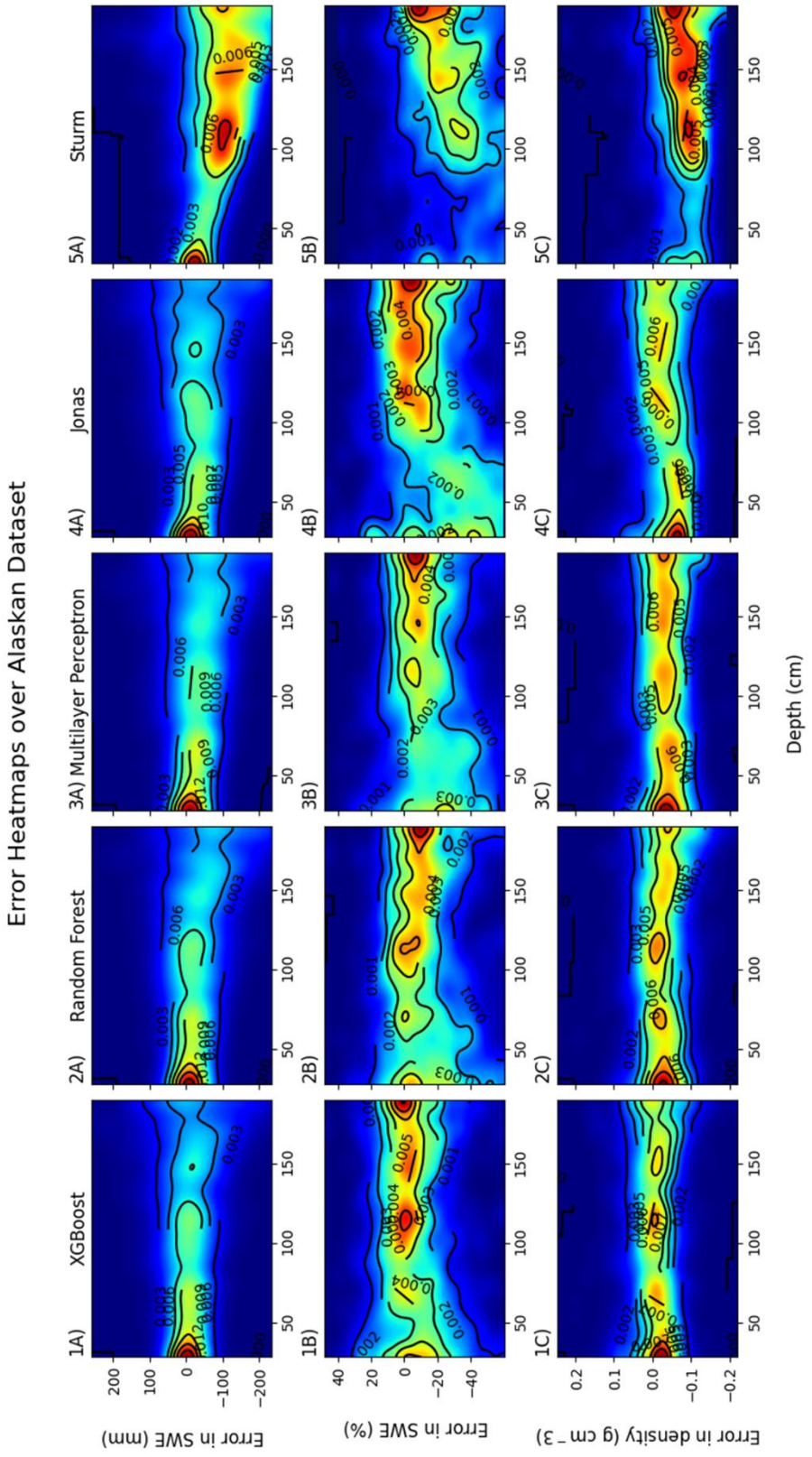

Figure 4.11: Heat maps of model performance for all the models (1-5) on Alaskan measurements. Better performance will produce narrower bands of high error densities centered around zero. An equal number of measurements exists for each 10 cm snow depth interval. Error distributions shown are a) SWE error in mm, b) relative SWE error in percent, and c) density error in g cm^{-3}. The error densities are displayed as a function of snow depth in cm. The scale goes from deep blue where there is low density of prediction errors to dark red for the highest error densities.

The MLP model performance can be seen in Figure 4.11-3. How the errors progress with increasing depth are similar to the Jonas model, but with better precision and accuracy. The error shows less spread with values centered around −10 mm at low snow depth to −40 mm at snow depth > 100 cm (a). The relative error at low snow depth is less probable below 40% with higher snow depth having an error of [−20%, 10%] (b). The density error at low snow depth is centered around −0.03 g/cm^3 (c).

The RF model error distributions seen in Figure 4.11-3 are very similar to those of the MLP. The SWE error (a) and the relative SWE error (b) look like the MLP model's but with higher precision. The better precision is best seen when comparing the relative SWE at low snow depth where the RF model has a higher density of errors in the range [−20%, 10%] at snow depth ≈ 75 cm. The density error probability of the RF model is centered around 0.025 g/cm^3 at low snow depth and produces narrower distribution bands than the preceding models (c).

The XGB model shows the best error distributions seen in Figure 4.11-1. The SWE error probability is centered around 0 mm for low snow depth spreading out as snow depth increases (a). The relative SWE error probability shows that the XGB model has the highest accuracy for predictions with snow depth < 100 cm. The relative SWE error ranges from [−20%, 15%] (b). The density error probability is centered around 0.02 g/cm^3 at low snow depth and around 0.00 g/cm^3 for snow depth > 100 cm (c).

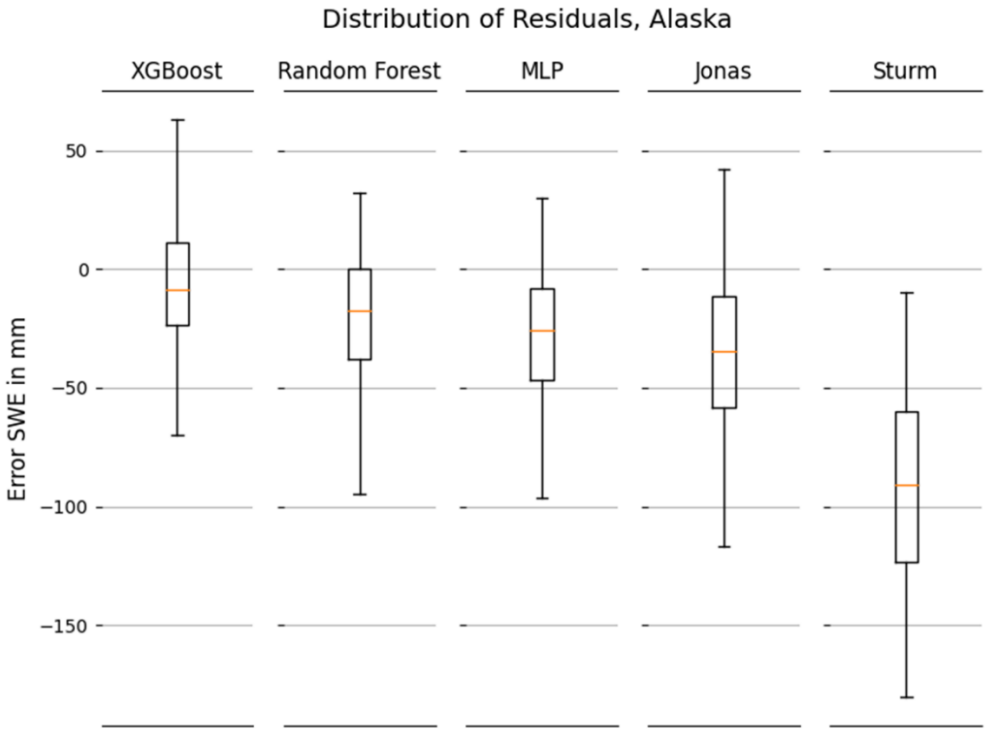

Figure 4.12: Box and whisker plot comparing the distribution of error residuals for the Alaskan validation. The whiskers mark the 5th and 95th percentile while the box shows the 25th and 75th. The median of the residuals is shown with a red line. The Jonas and Sturm models have been included as their results are comparable with these models.

Figure 4.12 shows the distribution of the error residuals for all five models. The regression models are included as they are more comparable in performance when applied to the Alaskan test dataset. For the

XGB model, 90% of the residuals are found within a range of [−70.0 mm, 63.2 mm] and 50% within [−23.8 mm, 11.4 mm]. The median value of the residuals is −8.5 mm. For the RF model 90% of the residuals are within [−94.8 mm, 32.0 mm] and 50% within [−37.8 mm, 0.1 mm] with a median prediction error of −17.0 mm. The residual distributions for the MLP model are 90% within [−96.7 mm, 29.8 mm] and [−46.8 mm, −8.0 mm] with a mean of −25.8 mm. Here it can be seen that all the best performing models in section 4.1 to some degree underestimate the SWE in Alaska. Of these the XGB model shows the lowest negative bias, while the MLP model shows the highest. The regression models can be seen to follow behind in terms of accuracy with the Jonas model performing best with 90% of the residuals within [−116.7 mm, 41.9 mm] and 50% within [−58.1 mm, −11.7 mm] and a median of −34.9 mm. The Sturm model has the largest spread of the residuals with 90% of the residuals within [−180.0 mm, −9.7 mm] and 50% within [−123.3 mm, −59.8 mm] and a median of −91.0 mm.

4.2.3 Further cross-validation

Table 4.7: The best number of estimators and the corresponding best R^2 score for the US states used for cross-validation.

State	N Estimators	R^2
Alaska	24	0.929
California	325.0	0.927
Idaho	100.0	0.956
Montana	75.0	0.963
Nevada	250	0.941
New Mexico	125	0.936
Oregon	75	0.946
South Dakota	650	0.845
Utah	375	0.931
Washington	50	0.936
Wyoming	25	0.949

The optimal parameters for the XGB model differed between section 4.1 and section 4.2. The parameters that gave the best results for a block bootstrap dataset were found to be those in section 4.2. Therefore, the same block bootstrap validation was performed on other US states to find their best parameters and observe if, or how, they differ. In these validations the number of layers was kept static at 7 while the number of estimators ranges from [1, 1000], testing every 25 for a total of 41 runs. The dataset used for training is the USCN training dataset with the stations from the validation state removed. The same state's measurements from the USCN validation and test datasets were also added to the validation. The results can be seen in Figure 4.13 and show that when the different states' datasets have different parameters where the model has the best performance. The best numbers of estimators yielding the best results as well as the corresponding R^2 score are found in Table 4.7. These parameters are not necessarily the best for each states' dataset, but it illustrates how overfitting is affecting areas differently. Some regions such

as Wyoming and Washington have few estimators for the best results, like Alaska. Other areas, like California and Utah, have better results with significantly more estimators.

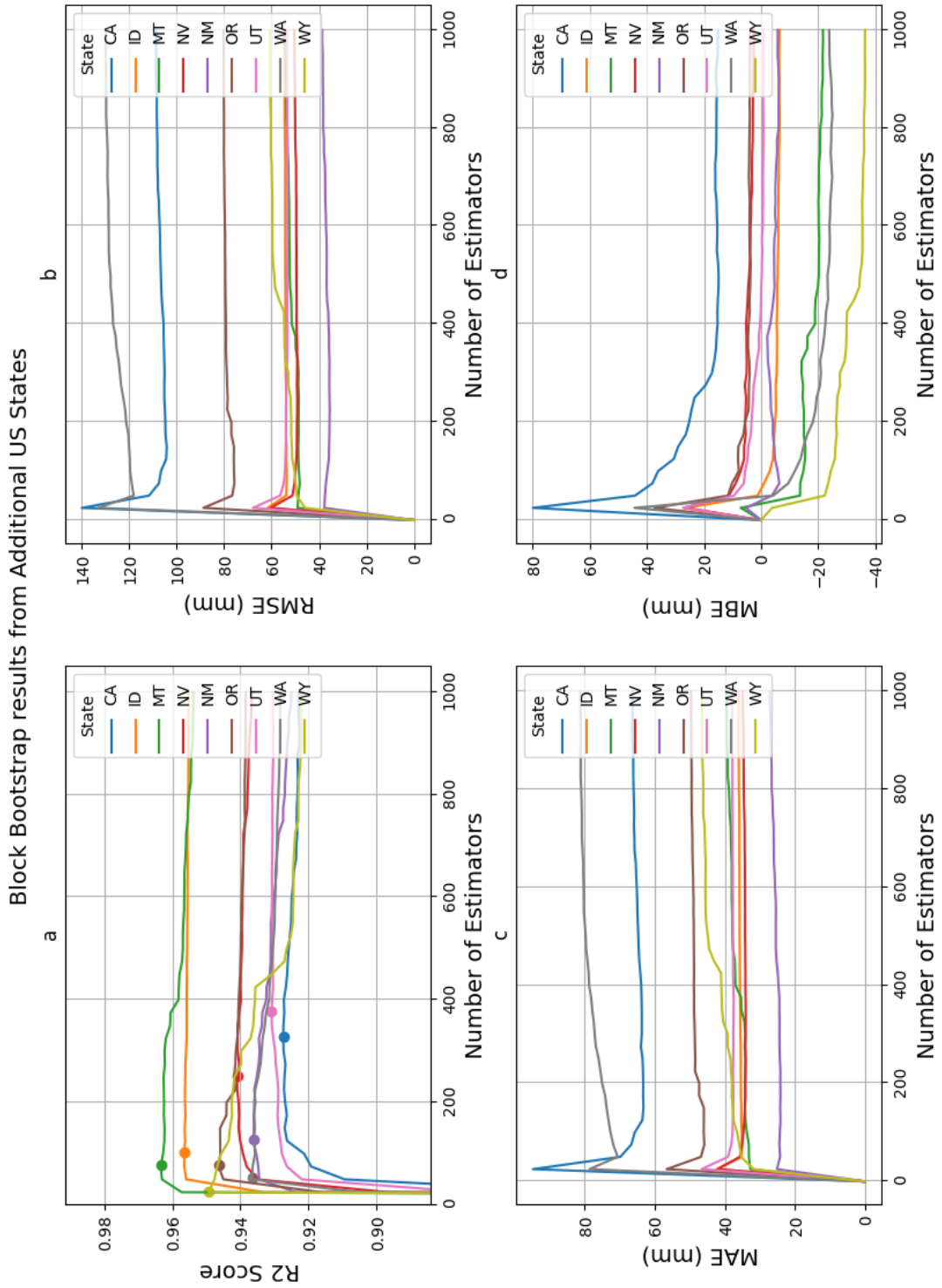

Figure 4.13: Cross validation performed on different subsets of the SNOTEL dataset part of the USCN dataset. The measured metrics are a) R^2 score with the highest point marked with a sphere, b) the RMSE, c) the MAE, and d) the MBE.

Using the XGB model with the Alaskan test and validation datasets for training and validation respectively, a grid search was performed to investigate the performance when using only regional measurements to construct the model. The grid search uses layers in the range [2, 11] and number of estimators [10, 200]. This is still a block bootstrap method as the sites validated for are not used in training. The results are shown in Figure 4.14 and show different parameters to give the best results than those found in section 4.2.1. The best parameters were 3 layers and 200 estimators. The best results from using these parameters for the validation gave an R^2 score of 0.943, a 1.6% improvement to the results from section 4.2.2. Not as great results was found when doing the same test for the RF model, which had a best R^2 score of 0.92, a 0.06% increase in performance. None the less it shows that as good results can be produced with a much smaller dataset. Fitting the Jonas model to the Alaskan validation dataset resulted in an R^2 score of 0.91, a considerable increase to the 0.731 when trained on USCN and applied to the same test set.

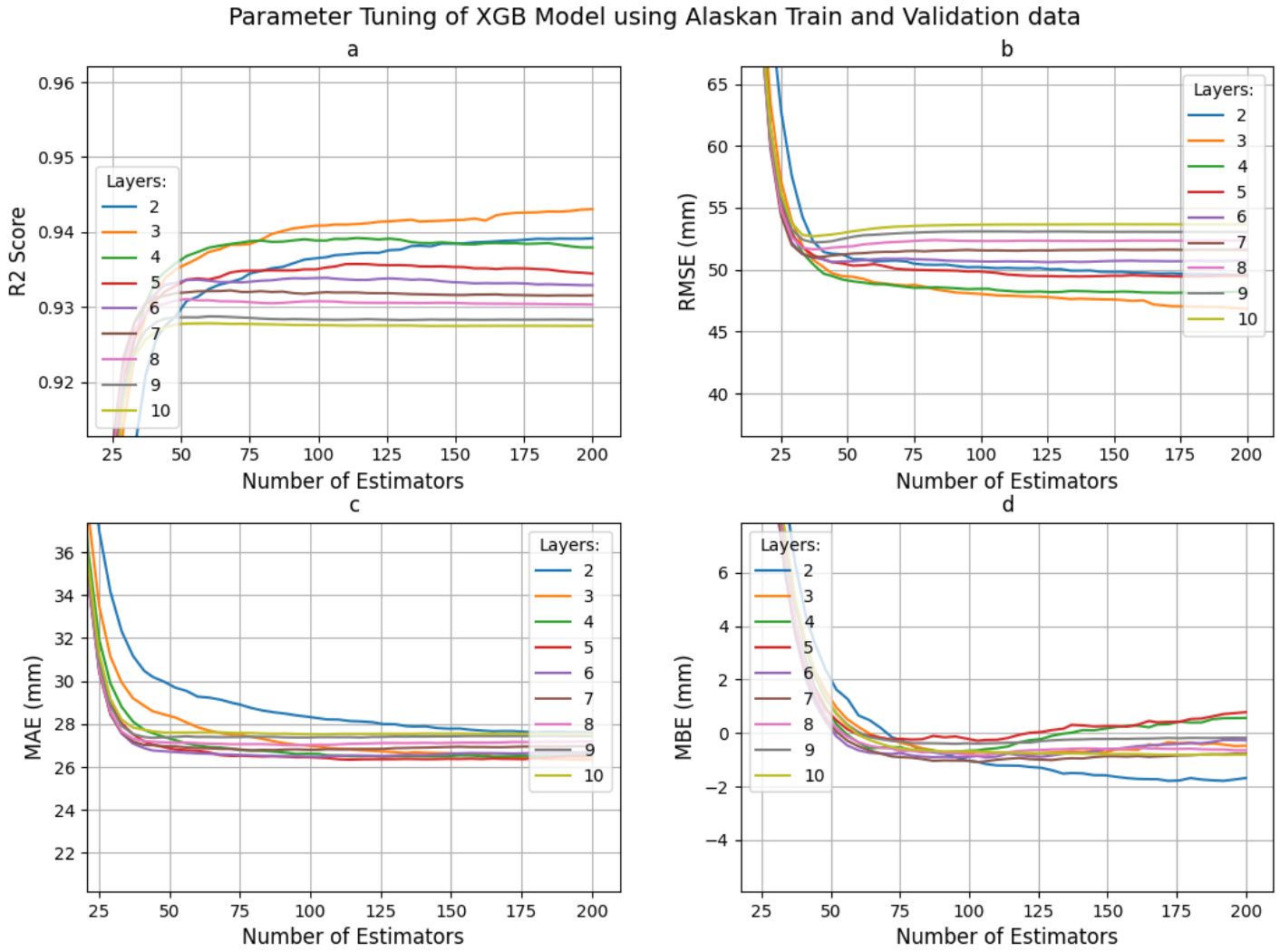

Figure 4.14: Grid search showing XGB model performance when using the Alaskan test set as the training dataset and validating against the Alaskan validation dataset. The metrics plotted are a) R^2 score, b) RMSE (mm), c) MAE (mm), and MBE (mm).

Table 4.8: The frequency and gain score for the XGB model constructed in section 4.1.2 and 4.2.2 and the gain feature score for the RF model constructed in section 4.1.2 and 4.2.2

Model	XGBoost			XGBoost			Random Forest		
Feature importance metric	Gain			Frequency			Gain		
Parameter | Layer/number of estimators	7 / 24	9 / 1500	Differance	7 / 24	9 / 1500	Differance	12 / 60	30 / 300	Differance
Snow depth	89.13 %	85.84 %	-3.69 %	19.61 %	15.23 %	-22.35 %	94.22 %	92.62 %	-1.69 %
Elevation	0.34 %	0.88 %	156.14 %	9.76 %	9.97 %	2.20 %	0.29 %	0.46 %	58.57 %
Day of year	3.22 %	3.91 %	21.21 %	9.76 %	6.63 %	-32.06 %	1.71 %	1.81 %	5.36 %
Days without snow	0.33 %	0.30 %	-9.85 %	1.28 %	6.76 %	429.67 %	0.06 %	0.18 %	180.53 %
Number of frost-defrost cycles	0.70 %	0.45 %	-35.49 %	3.03 %	7.87 %	159.32 %	0.13 %	0.28 %	108.67 %
Accumulated positive degrees	0.21 %	0.50 %	134.26 %	4.55 %	9.50 %	108.70 %	0.13 %	0.31 %	140.82 %
Average age of snow cover	0.97 %	1.04 %	7.16 %	8.10 %	7.42 %	-8.46 %	0.52 %	0.67 %	28.45 %
Number of layer	0.40 %	0.45 %	11.07 %	1.38 %	6.95 %	403.72 %	0.09 %	0.22 %	136.62 %
Accumulated solid precipitation	0.42 %	1.22 %	187.24 %	20.65 %	8.06 %	-60.97 %	0.53 %	0.73 %	37.93 %
Accumulated solid precipitation (last 10 da	0.45 %	0.61 %	36.96 %	7.00 %	6.63 %	-5.22 %	0.16 %	0.30 %	91.00 %
Total precipitation last 10 days	0.28 %	0.36 %	26.40 %	1.79 %	5.80 %	223.57 %	0.10 %	0.23 %	128.88 %
Average temperature last 6 days	3.18 %	3.50 %	9.90 %	10.48 %	7.55 %	-27.91 %	2.00 %	2.11 %	5.40 %
Snow class	0.35 %	0.95 %	170.60 %	2.62 %	1.65 %	-37.13 %	0.06 %	0.10 %	66.22 %

4.3 Feature Importance

Looking at the feature importance metrics can give some understanding of the inner workings through which these models arrive at their predictions. For the XGB model, the two feature importance scores looked at here are gain and frequency. They were obtained from the XGB model created in section 4.1.1 and 4.2.1. The feature fain score was obtained from the RF model from the same sections. Gain is a metric for how much relative contribution each explanatory variable has on each tree. In table 4.8 the feature importance metrics are shown with the gain is normalized and presented as the percentage. The percentage difference in the feature importance is also calculated. The frequency is the percentage of times a variable appears in the model as a split node criterion. For both the XGB and RF models the snow depth is the most important feature. This is perhaps not surprising, as it is the feature that most directly describes the local hydrology. The gain shows that the snow depth is almost one order of magnitude above that of the second most important feature, day of year.

For the XGB model, when validated over the Alaskan dataset, the variable with the highest frequency is the accumulated solid precipitation. This is one of the variables calculated from ERA5-Land data. This feature has slightly higher frequency than snow depth, while the gain is still amongst the lowest scoring. This changes for the more complex version where the variable has the fourth highest gain. Apart from the snow depth, the day of year and average temperature in the last six days are the variables with the highest gains. This is true for both parameter combinations and for both models. The day of year has for a long time been known to have predictive power and is used in the regression models (Sturm *et al.* 2009). Several variables have a low gain score, but the days without snow and total precipitation in the last 10 days are consistently low scoring between iterations and models. The difference in the feature importance could give some insight as to what variables rises in importance when the models are trained. Accumulated precipitation is the variable that has the highest gain increase for the XGB model. This variable is also one of the higher scoring ones for the RF model.

Further feature importance metrics were obtained from the XGB model to investigate the progression. The XGB model was chosen for its good performance and since its precision rose and declined as the number of estimators increased. This was done by getting the feature importance metrics from 100 runs with the number of layers being 7 and the number of estimators in the range [1, 100]. The results are seen in Figure 4.15. Beginning at the number of estimators being 1, as more are added, the frequency of the accumulated solid precipitation increases. It then starts decreasing in importance at around 30 estimators, coinciding with the Alaskan R^2 score declining. The average temperature in the last 6 days becomes the most frequently used variable at around 55 estimators. As more estimators are added, the snow depth variable rises in frequency and eventually becomes the most frequently used feature as seen in Table 4.8

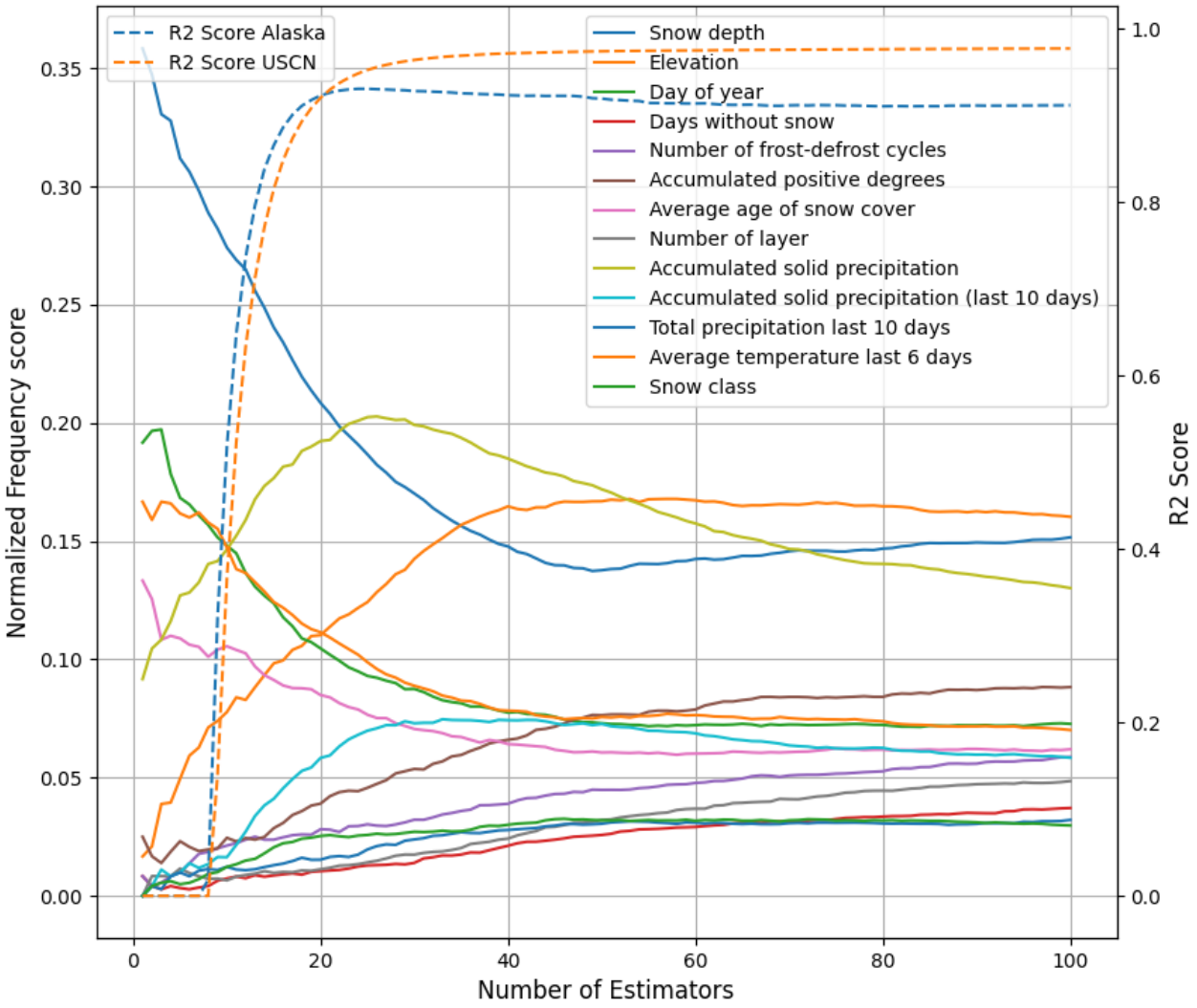

Figure 4.15: Feature progression as a function of number of estimators, the scores have been normalized to fit alongside the corresponding R^2 score calculated for both the Alaskan and USCN

4.4 Computational cost

Table 4.8: Model execution speed for models trained in section 4.1.2.

Model	Training Duration (s)
XGBoost	147
RF	400
MLP	72120
Jonas	6.1
Sturm	1.1

Execution time is another factor of interest when building models, and for the models trained in this paper it varied widely. The regression models are far less complex and thus computationally less costly, with processing time for a complete execution measured in seconds. Following those are the CART models, both having run times that can be measured in minutes. Of the CART models the XBG model turned out to be the faster one. Lastly comes the MLP ensemble model with a considerably long execution time, taking over 20 hours to fully train. This stems from the 20 ensemble members being trained individually and a new model being constructed for each snow class. For the best performance it was found in section 4.1.1.3 that many hidden layers and a large number of epochs were needed, further adding to the execution time. The run time for training each of the best models, as described in section 4.1.1, are shown in Table 4.8 and the equipment used is shown in appendix Table A1.

5. Discussion

The regression models in this paper have considerably higher accuracy than in previous studies (Sturm *et al.,* 2010; Ntokas *et al.,* 2021). This is due to the removal of faulty measurements of the datasets. The regression models are especially sensitive to erroneous snow depth measurements due to it being an exponent in the equations. The inclusion of ERA5-Land data to the more complex models could give them the ability to correct for false data and could be an explanation as to why less improvements were seen in the MLP model compared with Ntokas *et al*. (2021). In terms of their performance, high generalization errors were found for both the Jonas and Sturm models. They had good accuracy scores for the areas where they were trained, with both achieving R^2 scores greater than 0.94 for the USCN test dataset. The regression models were found to be best suited when trained with regional data.

The MLP model was reexplored through the hidden layer and epoch parameters. Ntokas *et al.* (2021) noted that their model had a bias to underestimating SWE for SWE measurements above 2500 mm. The MLP model produced in this paper was not found to harbor this upper boundary when tested on the USCN dataset. This could be a result of the extended dataset containing larger numbers of deeper SWE measurements, or the far greater number of layers and epochs used. Most likely a combination of these led to the improved reliability of the MLP model in deeper snow when not using block bootstrapping. When using block bootstrapping, the number of hidden layers needed to be drastically reduced to minimize the generalization error. The initial findings showed a generalization error of 57.4 mm. Reducing the number of hidden layers from in the hundreds down to only 2 reduced the generalization error to 0.3. The final resulting R^2 score of 0.894 improved greatly upon the initial score of 0.748. In development, the use of a single MLP model proved better suited than one for each snow class. This was because there simply existed too few measurements of some snow classes in Alaska.

The RF model showed the lowest generalization error over all the tests with 15.57 mm when testing on the Alaskan dataset with the USCN validated model. This could be due to the internal bootstrapping feature of the model, where the training dataset is used portion by portion when constructing the model. The number of layers in the RF model was the most deciding parameter regarding whether the model overfitted or not. The way the RF model uses additional estimators is by creating more trees uniquely, meaning that the trees do not influence each other. This is in contrast to the XBG model where the trees are predicting the previous residuals. The difference in construction becomes evident when plotting performance scores as a function of the number of estimators. The RF model shows differing performance mostly based on the number of layers while the XGB gets better performance as a function of both layers and estimators.

In regard to bootstrapping method, using a randomly selected subset was found to negatively affect the generality of the non-regression models. The findings were that higher complexity increased model performance. This reduces the usefulness of the three-way train, validation, and test split, as the subsets will perform better with the same parameters. The XGB model and RF model both showed similar performance for this validation with R^2 scores of 0.985. The MLP model achieved an R^2 score of 0.980. The models trained in this fashion was found to be overfitted to various degrees as they all had considerable generalization errors when tested against Alaskan data. The extremely high accuracy scores

are therefore of little value when not estimating for these areas. This is true to a lesser extent for the RF model. The measurements validated against were not seen by the model in training, but they are close in both temporal and spatial dimensions. The added complexity therefore likely overfits to learn the individual behaviors of the stations and regions.

When applying a block bootstrap approach, much simpler models produced the best results. When validated over Alaskan data, the RF model outperformed the MLP model while both were outperformed by the XGB model. The RF model's R^2 score of 0.916 is 1.2% lower than that of the XGB model, and the MLP model's R^2 score of 0.894 is 3.6% lower than the XGB. The XGB model, with an R^2 score of 0.934, seems to generalize the variable-SWE relationship best, and would be the preferable method when modeling areas with few previous SWE measurements. Iterating the XGB model further would yield results more like those of the RF and MLP models with a similar bias. The way the XGB model trains iteratively explains why it will overfit when adding too many estimators. This can also explain the downwards trending bias if the model starts off with a large bias and is minimizing it with each added estimator. Since the best performance was found during this downwards bias trend, it could be a combination of a somewhat well-trained model and early stopping acting as a bias offset. However, even though the model bias being equal to the regional bias might be the explanation, cross-validation of the findings found that increasing the number of estimators will at some point result in larger generalization error.

It is impossible to say for certain exactly what causes the models to overfit, but some properties of snow could help explain it. The factors identified by Zhong *et al.* (2021) causing local variations in density, for example vegetation and topography, are not captured in the model features. They could contribute to give each station a unique snow depth-SWE relationship. It is not only station specific characteristics that are being overfitted. Regions have different parameters where the XGB model gives the best results. This suggests that some of the fitting done is more general, and some areas will benefit more from this than others. At around 400 estimators the XGB model seems to be overfitting for most areas and produce deteriorating results. The use of early stopping can therefore be beneficial when estimating SWE using the XGB model. It should be kept in mind that all other fitting parameters were kept static, and the point at which the model starts overfitting is a function of all parameters.

When constructing the XGB model using only local data, the results were better than those found when using foreign training data and local validation. This suggests that fewer regional measurements have higher predictive power than a larger dataset obtained from elsewhere. The benefits of a small local dataset were also found when using the RF model, although it was not found to be an improvement to using foreign data. When using only local training data, the XBG model had the best results with only 3 layers, as opposed to 7 layers when using it as validation for foreign training.

The best performing XGB parameters when validating for a given area seem to be coupled with the parameters that give the lowest bias. The best performing RF and MLP models on the other hand both had significant negative biases and the regression models even greater. This could be caused by regional differences in snow densification. The Alaskan dataset was found to have a lower mean density than the USCN. Regional bias could also explain why, when validating over different areas, different parameter

combinations yield the best results. The XGB cross validation shows that the MBE converges on different values for different areas.

There are several points to consider when choosing which model and approach used to estimate the SWE, chiefly measurement availability. If the model is to be applied to an area where no measurements exist, and with no knowledge about the regional bias, the RF model would be the best candidate. This is due to the necessity of using training data from outside the region and lack of ability to validate. The RF model was found to have the lowest generalization error for this scenario in section 4.1.2. The downside of this approach is that the bias of the model is still unknown. This could be remedied by having only a few measurements from the relevant region. The best approach in this case, comparable to section 4.2, would be to use these measurements as a validation for an XGB model with training data from elsewhere. The XGB model proved best when using an early stopping scheme, but this requires some existing regional knowledge. If the region should have enough measurements available from enough sites to train and validate with these, then it was found in section 4.2.3 that this gives the best results. What the dataset size threshold for this to the preferred method is, as well as the performance impact dataset size has on the model performance, is not explored in this paper.

Results presented here strengthen the case for use of reanalyzed data when creating SWE models. The accumulated solid precipitation calculated from ERA5-Land data was the leading feature in terms of frequency where the XGB model was best at generalizing the SWE predictions. The average temperature in the last 6 days had the highest gain score of the ERA5-Land derived variables for both the RF and XGB model. Its importance to the model predictions was similar to that of the day of year. 10 of the model's 13 explanatory variables came from ERA5-Land data, and they all participate in the model prediction to some degree. The ERA5-Land derived features make the models easy to apply on new snow depth measurements as only date, location, and snow depth is needed to produce the variables that go into the models.

The importance of snow classes from Sturm *et al.* (1995) proved less important in the feature analysis of the CART models. This feature had the lowest importance in terms of gain for both XGB and RF models when validating for USCN and was among the lower scoring features for the Alaskan validated models. This should not be interpreted as the snow classes having low predictive power since: (a) The findings in section 4.1.2 are highly overfitted and (b) the snow class classifications stem from many of the same variables also supplied to the model. This would mean that more complex models could internally generalize these snow classifications to some degree using the Odry *et al.* (2020) explanatory variables.

6. Conclusion

The new machine learning approaches investigated in this paper offer several advantages over previous models. First and foremost is the improved reliability of the predictions, with the best XGB model having a 3.6% better performance than the MLP model when tested on Alaskan test data. The RF model also had a better performance than the MLP model, while its true strength lies in the low generalization error. Secondly comes the lower computational cost with the time needed for training an XGB model being 0.2% of that of the MLP ensemble. Advantages with shorter computational times are numerous as parameters can be optimized faster and results can quickly be compared. Lastly comes the inherent complexity, this gives the RF and XGB models the ability to take parameters such as snow class as a direct column feature. The CART models also have a way of handling missing data (Chen & Guestrin (2016); Breiman (2001)). This is an advantage as data points can still be evaluated without all features being included.

The RF model was found to be the preferred model in areas where no SWE measurements exist, while the XGB model is the preferred option if some are available for validation. The most advantageous method, where enough measurements exist, is training an XGB model with the available SWE measurements. It is very important to find the correct parameters as all the models investigated were found to be prone to overfitting. Parameters producing simpler models are preferred as this leaves less opportunity for the model to memorize individual station characteristics.

The findings from using a station-based block bootstrap approach brought up some concerns regarding the use of machine learning algorithms for SWE predictions. It was found that the XGB, RF, and MLP models can easily be overfitted, leading to unreliable results when tasked with predicting SWE from sites they have not previously seen. When training, the models use the training data in several iterations and will inevitably start fitting results to these exact measurements. Therefore, it is important to be aware of the goals of the model when selecting parameters. The best approach found to modeling SWE in regions where some measurements exist are to train the models with these and excluding outside data. For the XGB model this approach resulted in an R^2 score of 0.943, a 1.6% improvement in performance compared with using the measurements as validation for foreign training data. What the performance impact of a growing dataset has on model performance could be an area for future research.

Future research could also focus on the use of different bootstrapping schemes. A block bootstrap approach could be explored in the temporal dimension as well, as measurements obtained close in time from the same site might be very identical and cause overfitting. Assessing and possibly omitting measurements that are too similar might prompt a more general understanding of the variable-SWE relationship to be found.

7. Sources

Anderson, J., and Wirt J. (2008). Ultrasonic snow depth sensor accuracy, reliability and performance, paper presented at 76th Western Snow Conference, pp. 99–105.

ASM (2021). How XGBoost Works. Amazon Web Services, Amazon SageMaker Developer Guide, pp. 1960-1961.

AML (2021). Model Fit: Underfitting vs. Overfitting. Amazon Web Services, Amazon Machine Learning Developer Guide, pp. 17-18.

Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R., and Dozier, J. (2006), Mountain hydrology of the western United States, Water Resour. Res., 42, W08432, doi:10.1029/2005WR004387.

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Bruland O., Færevåg, Å., Steinsland, !., Liston, G. E., Sand, K. (2015) Weather SDM: estimating snow density with high precision using snow depth and local climate. Hydrology Research 1 August 2015; 46 (4): 494–506. doi: https://doi.org/10.2166/nh.2015.059

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O. Niculae, V., Prettenhofer, P., Gramfort, A. Grobler, J. Layton, R., Vanderplas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project

Chen, T. and Guestrin, C. XGBoost (2016). A scalable tree boosting system, Association for Computing Machinery (2016), pp. 785-794

Chicco D., Warrens M. J., Jurman G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science* 7:e623 https://doi.org/10.7717/peerj-cs.623

Cook, J., et al. (2013). "Quantifying the consensus on anthropogenic global warming in the scientific literature," Environmental Research Letters Vol. 8 No. 2, (15 May 2013); DOI:10.1088/1748-9326/8/2/024024

Davis, R. (1973), Operational snow sensors, paper presented at 30th Eastern Snow Conference, pp. 57-70

Deems, J., Painter, T., and Finnegan, D. (2013). Lidar measurement of snow depth: A review. *Journal of Glaciology, 59*(215), 467-479. doi:10.3189/2013JoG12J154

Demaria, E. M. C., Roundy, J. K., Wi, S., and Palmer, R. N. (2016). The Effects of Climate Change on Seasonal Snowpack and the Hydrology of the Northeastern and Upper Midwest United States, *Journal of Climate*, *29*(18), 6527-6541. https://journals.ametsoc.org/view/journals/clim/29/18/jcli-d-15-0632.1.xml

Dettinger, M. D., and Cayan D. R. (1995). Large-Scale Atmospheric Forcing of Recent Trends toward Early Snowmelt Runoff in California". Journal of Climate 8.3 (1995): 606-623. https://doi.org/10.1175/1520-0442(1995)008<0606:LSAFOR>2.0.CO;2.

Essery, R. (1997). Seasonal snow cover and climate change in the Hadley Centre GCM. *Annals of Glaciology, 25*, 362-366. doi:10.3189/S0260305500014282

Fierz, C., Armstrong, R. L., Durand, Y., Etchevers, P., Greene, E., Mcclung D., Nishimura K., Satyawali, P., and Sokratov, S. (2009). The international classification for seasonal snow on the ground (UNESCO, IHP (International Hydrological Programme)–VII, Technical Documents in Hydrology, No 83; IACS (International Association of Cryospheric Sciences) contribution No 1).

Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. Q J R Meteorol Soc. 2020; 146: 1999– 2049. https://doi.org/10.1002/qj.3803

Jakubovitz, D., Giryes, R., and Rodrigues, M. R. D. (2019). Generalization Error in Deep Learning.

Jennings, K.S., Winchell, T.S., Livneh, B. et al. (2018). Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere. Nat Commun 9, 1148 (2018). https://doi.org/10.1038/s41467-018-03629-7

Johnson, J. B. (2004). A theory of pressure sensor performance in snow. Hydrol. Process., 18: 53-64. https://doi.org/10.1002/hyp.1310

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

Kinar, N. J., and Pomeroy, J. W. (2015). Measurement of the physical properties of the snowpack, *Rev. Geophys.*, 53, 481– 544. doi:10.1002/2015RG000481.

Lange, H.. and Sippel, S. (2020). Machine Learning Applications in Hydrology. 10.1007/978-3-030-26086-6_10.

Liaw, A. and Wiener, M. (2001). Classification and Regression by RandomForest. Forest. 23.

Magnusson, J., Nævdal, G., Matt, F., Burkhart J. F., and Adam Winstral (2020). Improving hydropower inflow forecasts by assimilating snow data. Hydrology Research 1 April 2020; 51 (2): 226–237. doi: https://doi.org/10.2166/nh.2020.025

Mote, P.W., Li, S., Lettenmaier, D.P. et al. (2018) Dramatic declines in snowpack in the western US. npj Clim Atmos Sci 1, 2 (2018). https://doi.org/10.1038/s41612-018-0012-1

Murray, G. and Scime, A. (2010). Microtargeting and Electorate Segmentation: Data Mining the American National Election Studies. Journal of Political Marketing. 9. 143-166. 10.1080/15377857.2010.497732.

Ntokas, K. F. F., Odry, J., Boucher, M.-A., and Garnaud, C.: Investigating ANN architectures and training to estimate snow water equivalent from snow depth, Hydrol. Earth Syst. Sci., 25, 3017–3040, https://doi.org/10.5194/hess-25-3017-2021, 2021.

Odry, J., Boucher, M. A., Cantet, P., Lachance-Cloutier, S., Turcotte, R., and St-Louis, P. Y.: Using artificial neural networks to estimate snow water equivalent from snow depth, Can. Water Resour. J., 45, 252–268, https://doi.org/10.1080/07011784.2020.1796817, 2020.

Prowse, T. D., Furgal, C., Chouinard, R., Melling, H., Milburn, D., & Smith, S. L. (2009). Implications of Climate Change for Economic Development in Northern Canada: Energy, Resource, and Transportation Sectors. Ambio, 38(5), 272–281. http://www.jstor.org/stable/25515855

Qi J., Du J., Siniscalchi S. M., Ma X. and Lee C. -H., (2020). "On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression," in *IEEE Signal Processing Letters*, vol. 27, pp. 1485-1489, 2020, doi: 10.1109/LSP.2020.3016837

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J., Thériault, J. M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K., & Gutmann, E. (2012). How Well Are We Measuring Snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed, Bulletin of the American Meteorological Society, 93(6), 811-829. Retrieved May 5, 2022, from https://journals.ametsoc.org/view/journals/bams/93/6/bams-d-11-00052.1.xml

Roulston, M. S., & Smith, L. A. (2002). Evaluating Probabilistic Forecasts Using Information Theory, Monthly Weather Review, 130(6), 1653-1660. Retrieved Mar 5, 2022, from https://journals.ametsoc.org/view/journals/mwre/130/6/1520-0493_2002_130_1653_epfuit_2.0.co_2.xml

Sabater, M. J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Skaugen, T. (1998). Estimating the mean areal snow water equivalent from satellite images and snow pillows. https://hdl.handle.net/11250/2830976

Snauffer, A. M., Hsieh, W. W., Cannon, A. J., and Schnorbus, M. A. (2018). Improving gridded snow water equivalent products in British Columbia, Canada: multi-source data fusion by neural network models, The Cryosphere, 12, 891–905, https://doi.org/10.5194/tc-12-891-2018, 2018.

Stewart, I., Cayan, D., and Dettinger, M. (2004). Changes in snowmelt runoff timing in western North America under abusiness as usual'climate change scenario. Climatic Change. 62. 217-232. 10.1023/B:CLIM.0000013702.22656.e8.

Sturm, M., and Liston, G. E. (2021). Revisiting the Global Seasonal Snow Classification: An Updated Dataset for Earth System Applications. Journal of Hydrometeorology 22, 11, 2917-2938. https://doi.org/10.1175/JHM-D-21-0070.1

Sturm, M., Holmgren, J., and Liston, G. E. (1995). A Seasonal Snow Cover Classification System for Local to Global Applications, *Journal of Climate*, *8*(5), 1261-1283.

Sturm, M., Taras, B., Liston, G. E., Derksen, C., Jonas, T., & Lea, J. (2010). Estimating Snow Water Equivalent Using Snow Depth Data and Climate Classes, Journal of Hydrometeorology, 11(6), 1380-1394

Tarek, M., Brissette, F. and Arsenault, R. (2020). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. Hydrology and Earth System Sciences. 24. 2527-2544. 10.5194/hess-24-2527-2020.

USDA (2022). Automated Snow Monitoring, United States Departement of Agriculture. Retrieved 14.05.2022 https://www.nrcs.usda.gov/wps/portal/wcc/home/aboutUs/monitoringPrograms/automatedSnowMonitoring/

Vincent, W. (2010) Microbial ecosystem responses to rapid climate change in the Arctic. *ISME J* **4,** 1087–1090 (2010). https://doi.org/10.1038/ismej.2010.108

Vionnet, V., Mortimer, C., Brady, M., Arnal, L., and Brown, R. (2021). Canadian historical Snow Water Equivalent dataset (CanSWE, 1928–2020), Earth Syst. Sci. Data, 13, 4603–4619. https://doi.org/10.5194/essd-13-4603-2021, 2021.

Willmott C. J. and Matsuura K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance

Xu, Y., Jones, A., and Rhoades, A. (2019). A quantitative method to decompose SWE differences between regional climate models and reanalysis datasets. Scientific reports, 9(1), 16520. https://doi.org/10.1038/s41598-019-52880-5

Zhong, X., Zhang, T., Su, H., Xiao, X., Wang, S., Hu, Y., Wang, H., Zheng, L., Zhang, W., Xu, M., and Wang, J. (2021). Impacts of landscape and climatic factors on snow cover in the Altai Mountains, China, Advances in Climate Change Research, Volume 12, Issue 1, 2021, Pages 95-107, ISSN 1674-9278, https://doi.org/10.1016/j.accre.2021.01.005.

Appendix

Table A1: Computer specifications

Component	Model
CPU	Intel i7 3770k 3.5GHz
GPU	Nvidia GeForce 980
RAM	Kingston DDR3 1600MHz 8gb x2
Motherboard	MSI A520M PRO