https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

Evaluating a hierarchy of bias correction methods for ERA5-Land SWE in northern Canada

Neha Kanda[1] and Christopher G. Fletcher[1]

[1]Deptartment of Geography & Environmental Management, University of Waterloo, Ontario, Canada

Correspondence: Christopher G. Fletcher (chris.fletcher@uwaterloo.ca)

Abstract. Precise estimates of Snow Water Equivalent (SWE) are crucial for informed decision-making in regions like Northern Canada, where snow cover significantly contributes to springtime discharge. However, the sparse nature of the existing SWE monitoring network poses a challenge to comprehensively understanding the SWE distribution and variability. Reanalysis products like ERA5-Land provide long-term continuous SWE estimates, but our evaluation identified a negative bias (-61mm) in the estimated SWE and maximum underestimation was observed at high elevation (>1500m) areas. To correct these biases, we applied four correction methods: Mean Bias Subtraction (MBS), Simple Linear Regression (SLR), Multiple Linear Regression (MLR), and Random Forest (RF). RF exhibited the highest performance, reducing the Root Mean Square Error (RMSE) by 78% and minimizing the annual mean bias from 61.2 mm to 0.01 mm. However, RF did not produce reliable SWE estimates for unseen spatial and temporal domains due to its limitation of not extrapolating beyond the training data.

1 Introduction

Snow cover plays a crucial role in shaping the hydrological cycle, particularly in high-latitude and high-elevation regions of the world, where it significantly contributes to springtime discharge(Kouki et al. 2023). Consequently, alterations in snow cover can lead to variations in water availability, impacting not only local ecosystems but also have significant implications for various human activities(Barnett et al. 2005). The Snow Water Equivalent (SWE), a key metric quantifying water content within snowpacks(Fierz et al. 2009), holds particular significance in regions north of 50 degrees latitude, where snow cover persists for a substantial part of the year. Thus, accurate measurements of SWE in snow-dominated Northern Canada are vital for understanding local climate dynamics and hold implications for effective water resource management(Shao et al. 2022).

Traditional methods for in-situ Snow Water Equivalent (SWE) measurements, like manual snow surveys, offer precise information but suffer from a lack of spatial representation, as they are confined to a single point (Meromy et al. 2012). Additionally, the drawbacks of manual snow surveys include their high cost, time-consuming nature, limited coverage in the northern latitudes of Canada (Fig1), and the fact that they provide only biweekly estimates(Brown et al. 2019). As a result, there is a heavy reliance on remotely sensed data which facilitates global coverage. However, Remotely sensed datasets like AMSR-E(Kelly 2009) and AMSR2(Imaoka et al. 2010; Tedesco and Jeyaratnam 2019), which, despite their widespread use, have limitations such as insufficient temporal coverage and coarse spatial resolution(Mortimer et al. 2020).

The state-of-the-art reanalysis dataset by the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5-Land,

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

provides a valuable decadal time series with multiple variables at 9 km spatial resolution, making it a popular dataset for climate studies(Muñoz-Sabater et al. 2021). However, ERA5-Land was found to produce biased estimates of SWE at high-elevation areas of North America(Kouki et al. 2023; Shao et al. 2022). This could be attributed to the spatial resolution of the dataset. In mountainous areas, the assumption of a smooth topography over a land grid of a few kilometers leads to an insufficient representation of the topography-induced effects and orographic precipitation(Chen et al. 2021). It was reported that the difference in elevation between station and ERA5-Land grid points resulted in biased estimates of temperature and precipitation over topographically diverse terrain like mountains(Zhao and He 2022; Velikou et al. 2022; Minola et al. 2020; Kouki et al. 2023). For such diverse terrains, even a grid size of 1 km may not adequately capture local site characteristics(Hastings et al. 1999). While large-scale gridded products may lack the fine details necessary to accurately depict local features, the combination of large-scale gridded data with finer details provided by additional covariates could offer a more comprehensive understanding of the spatiotemporal variations observed in manual snow surveys(Snauffer et al. 2018).

This fusion of gridded data with high-resolution covariates is possible through statistical and machine learning models. These models leverage the connection between SWE and environmental variables, termed predictors, to predict SWE values with higher accuracy(King et al. 2020; Snauffer et al. 2016; Shao et al. 2022). Since the relationship between SWE, geophysical, and climatic predictors is non-linear, non-linear models, like Random Forest, perform better than linear models due to their ability to capture complex relationships between SWE and environmental variables(King et al. 2020; Bennett et al. 2022). However, it was found that spatial prediction models especially Random Forest may not accurately extrapolate to unseen times and locations(Roberts et al. 2017) but none of the above-mentioned studies evaluated the extrapolation error of the best-performing model.

Currently, a comprehensive evaluation of biases and performance evaluation of different bias correction techniques over the entirety of northern Canada is lacking. Additionally, the capability of bias correction methods to extrapolate SWE information to unsampled locations/data-sparse regions like the Arctic remains unexplored. This research aims to address these gaps by:

1. Quantifying spatial and temporal biases in ERA5-Land SWE over Northern Canada.
2. Evaluating the performance of a hierarchy of bias correction techniques.
3. Assessing the capability of the best models to extrapolate SWE information to unsampled spatial and temporal domains.

2 Data and Methods

2.1 Field observations of snow

The Canadian Historical Snow Water Equivalent dataset (CanSWE) is a combined product of manual and automated SWE data from various locations across Canada(Vionnet et al. 2021). For this study, the data from approx 777 sites for the period of 1971-2019 was used for validation purposes. To ensure data integrity, sites with a time series length of less than one year were excluded from the analysis.

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

To quantify biases in ERA5-Land SWE, a point-to-pixel comparison was conducted, where the in-situ SWE values were matched with the nearest grid values from ERA5-L on matching dates. Outliers in both CanSWE and ERA5L were identified and removed using a Z-score threshold of 3.

To investigate the spatial transferability of the best prediction model, the study area was partitioned into 12 distinct ecoregions based on the classification given by Wiken (1986). The distribution of CanSWE sites across these ecoregions is depicted in Figure 1.

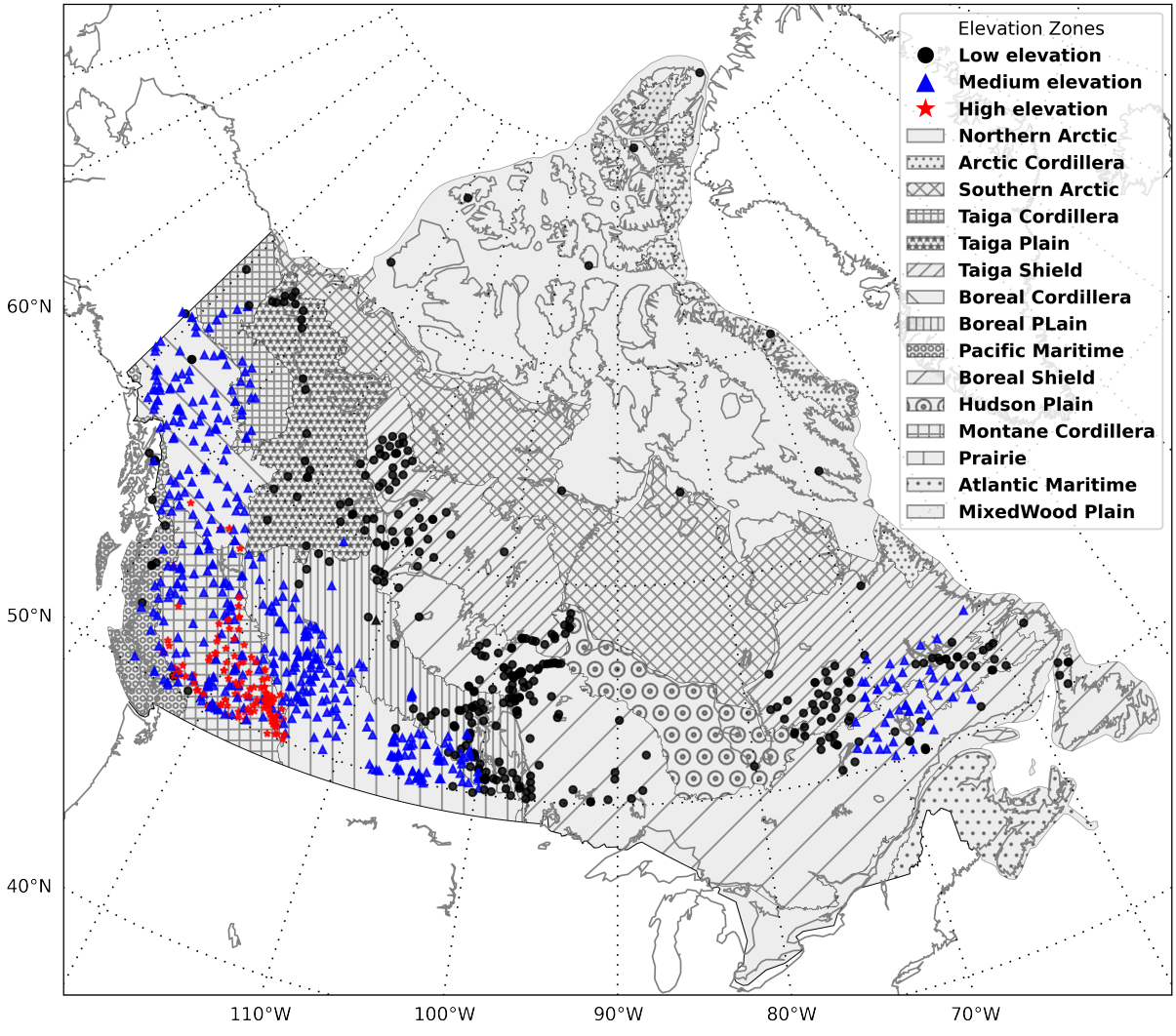

Figure 1. Spatial distribution of CanSWE sites denoted by points, categorized by elevation into Low (<500m), Mid (500-1500m), and High (>1500m). Shaded areas represent the Ecoregions of Canada.

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

2.2 Reanalysis Data

ERA5-Land, a fifth-generation global reanalysis dataset developed and released by the European Centre for Medium-Range Weather Forecasts (ECMWF), is the land only component produced at an enhanced spatial resolution of approximately 9 kilometers(Muñoz-Sabater et al. 2021). ERA5-Land runs without data assimilation and was generated without coupling with the atmospheric module making it computationally efficient(Muñoz-Sabater et al. 2021). These datasets encompass 50 different variables essential for characterizing the global dynamics of land-based water and energy cycles on an hourly basis.

For SWE, the snow depth (in m water equivalent) variable was used. For ERA5-Land grid elevation, geopotential was converted to geopotential height by dividing with gravitational acceleration,g(9.80665 m/s^2). Geopotential height is often referred to as orography on the surface of the earth and hence was used as grid elevation(Muñoz-Sabater et al. 2021). The data period of study was 1971-2019.

2.3 Site Characteristics/Ancillary datasets

Terrain features like elevation, Slope, and Aspect were derived from the Copernicus DEM (Glo 90) which is available at a spatial resolution of 90m(European Space Agency and Sinergise 2021). Here also, station locations were matched with the nearest grid values from Glo 90.

Elevation biases for station data were computed by subtracting the mean elevations of grid cells obtained from ERA5-Land geopotential height from those of the Digital Elevation Model (DEM). These elevation biases, specific to each product, were then utilized as predictors. To account for the impact of vegetation density on SWE, we used a normalized difference vegetation index (NDVI) as a proxy for vegetation density(Bennett et al. 2022). NDVI indicates the presence of active vegetation and helps assess factors like branch abundance and shrub height. We obtained maximum NDVI values from MOD13A1 MODIS/Terra product at a spatial resolution of 500 m(Didan et al. 2015)

To account for the temporal aspect of Snow Water Equivalent (SWE), we incorporated the fraction of the year that had elapsed at the time of observation as an additional predictor. This elapsed year fraction is represented using trigonometric predictors such as sine and cosine, given its periodic nature.(Snauffer et al. 2018).

2.4 Bias correction Models

We fit three different types of models to obtain bias-corrected SWE estimates. Mean Bias Subtraction (MBS) is the simplest bias correction technique which involves first calculating the average difference between ERA5-Land and in-situ SWE measurements. This calculated average difference is then subtracted from each ERA5LSWE estimate as shown in Equation 1.

$$\text{Mean Bias} = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i) \tag{1}$$

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

Table 1. Overview of predictors used in the study along with their data sources, and spatial resolutions.

S.No.	Predictor	Units	Source	Spatial Res (km)
1.	Snow water Equivalent	mm	ERA5-Land	9
2.	NDVI	unitless	MODIS	0.5
3.	Elevation	masl	Copernicus DEM	0.09
4.	Elevation Difference	m	ERA5L-DEM	-
5.	Slope	degrees	Copernicus DEM	0.09
6.	Aspect	degrees	Copernicus DEM	0.09
7.	Year	-	-	-
8.	Sine of 2π times the elapsed year fraction	-	-	-
9.	Cosine of 2π times the elapsed year fraction	-	-	-
10.	Latitude	degree decimals	-	-
11.	Longitude	degree decimals	-	-

, where N represents total number of observations.X_i depicts modeled SWE, Y_i depicts observed SWE and i represents an individual observation.

In a multiple linear regression model, the relationship between the dependent variable (Y) and multiple independent variables ($X_1, X_2, \ldots, X_k$) is represented by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{2}$$

where Y is the dependent variable, $X_1, X_2, \ldots, X_k$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for the respective independent variables, ε is the error term, capturing the unexplained variability in Y.

This equation describes a linear relationship between Y and multiple independent variables with an intercept (β_0) and coefficients ($\beta_1, \beta_2, \ldots, \beta_k$).

A random forest model is composed of numerous individual decision trees that operate collectively as an ensemble. Each tree within the random forest contributes its prediction by voting for a class (in classification) or providing a mean prediction (in regression). The final prediction is determined by the outcome with the highest number of votes(Liaw and Wiener 2002). Random forests offer several advantages, such as resilience to data correlations, a tendency to mitigate overfitting risks, and inherent feature randomization. However, a potential limitation is their complexity when utilized as a black-box system. In this study, we implemented random forests using the Scikit-learn package in Python, specifically employing the RandomForestRegressor class(Pedregosa et al. 2011). To address overfitting, model performance evaluation utilized a 5-fold cross-validation (5-fold CV) technique. The dataset was initially divided into 5 equal subsamples, with model training on 4 while using the remaining subsample for testing. This iterative process, repeated 5 times, ensured comprehensive testing across all subsamples, reducing overfitting risks.

However, selecting validation data randomly across the entire spatial area for cross-validation can introduce spatial autocorrelation, connecting training and validation data from nearby locations(Meyer et al. 2018). This can lead to overly positive

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

results, not accurately reflecting the model's true performance. These optimistic results hinder the estimation of the model's performance at locations not part of the training sample, known as the extrapolation error(Roberts et al. 2017).

To ensure a more independent assessment in cross-validation (CV), a common approach involves dividing the data into 'blocks' centered on specific points within the dependence structure, such as time or space(Roberts et al. 2017). Although 'block' cross-validations may show larger prediction errors compared to random cross-validation methods, they provide a closer estimation of the extrapolation error and are commonly used for assessing model transferability to new areas and times(Roberts et al. 2017). Additionally, selecting validation data randomly across the entire spatial area for cross-validation could lead to spatial auto-correlation, affecting result accuracy(Meyer et al. 2018).

To overcome this, we employed a spatial and temporal-based k-fold cross-validation approach. In Leave One Out CV (LOOCV), CanSWE sites were randomly divided into 5 groups, maintaining 90% as training sites and 10% as test sites. To assess model transferability to spatially distinct zones, the area was divided into 11 ecozones, training the model on 10 and testing on one. Temporal independence (LTOCV) was considered by using data from 1971-2014 as training and 2015-2019 as test data. Evaluation metrics, including Mean Absolute Error (MAE), coefficient of determination (R^2), and root mean squared error (RMSE), were computed on the model output within the test set.

Further, the Mean Decrease in Impurity (MDI) method was employed to calculate feature importance due to its computational efficiency(Bennett et al. 2022). In the Random Forest (RF) algorithm, the splitting rules aim to maximize impurity reduction resulting from a split. In assessing impurity importance, significance is attributed to splits causing a substantial decrease in impurity, and consequently, variables utilized for significant splits are deemed important. Following this rationale, the impurity importance for a variable Xi is determined by summing the impurity decrease measures across all nodes in the forest where a split on Xi occurred. This sum is then normalized by the number of trees in the forest(Ishwaran 2015).

Before model training, assessments for feature collinearity were conducted, particularly considering the impact of collinearity on Linear Regression models' feature importance. Variance inflation factors and pairwise correlation coefficients were computed to gauge feature collinearity and correlated features were removed.

The Random Forest (RF) model, implemented using the scikit-learn package in Python, underwent hyperparameter tuning to optimize model performance. The comprehensive predictor set underwent a 5-fold CV on randomly sampled data, culminating in selecting the best-performing model based on the highest R^2 and minimal RMSE. This optimal model was subsequently utilized for predictions on the test set.

3 Results

3.1 Quantification of bias in ERA5L

The study reveals a mean annual SWE bias and Root Mean Squared Error (RMSE) of approximately -61mm, and 223mm respectively, indicating notable discrepancies between observed and predicted SWE. These errors exhibit both temporal and spatial variability. To understand the temporal variability, the impact of seasons on SWE estimations is shown in Fig 2(a). On average, Spring (MAM), winter (DJF), and fall (ON) SWE are underestimated by approximately 87mm,42mm, and 13 mm

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

respectively.

However, this underestimation of SWE was not spatially consistent. To understand the spatial variability, ERA5L values were compared with CanSWE for three elevation zones: low elevation (<500m) (Fig 3a), mid-elevation (500-1500m)(Fig 3b), and high-elevation (>1500m)(Fig 3c). A total of 232, 415, and 133 sites are situated in low, mid, and high-elevation areas respectively (Fig1). High-elevation areas, characterized by deeper snowpacks (SWE values exceeding 400mm), experience substantial under-estimation by ERA5L (Fig 3c). This observation is further emphasized in the binned SWE values analysis in Fig 2(b), illustrating that ERA5L consistently underestimates SWE values exceeding 300mm. This underestimation could be attributed to the elevation difference between the ERA5L grid and the station elevation as we found that the sites with negative elevation bias depicted a negative SWE bias. Conversely, in low-elevation regions, ERA5L appears to overestimate SWE (Fig 3a) while it aligns most closely with ground SWE at mid-elevations (Fig 3b). Of the total 777 sites, the SWE was overestimated at 494 sites and underestimated at 283 sites with a relative mean bias of approximately 42%.

Figure 3 illustrates the SWE time series for both ERA5L and CanSWE at all elevation zones. There is an increasing SWE trend at low, mid-elevations, and overall study area, along with a corresponding decrease at high elevations as shown by both datasets. This highlights a robust alignment between datasets over different years, although SWE biases are evident in the time series. ERA5L captures the SWE trend at all elevation zones with a correlation of 0.98. The increasing trend in Fig 3 (b) and (d) is potentially a sampling issue. Post-1995, the inclusion of snow-pillow data (Fig A1) from high-elevation sites (Fig B2) seems to be a contributing factor, as the mean SWE from these sites is higher (Fig B3) compared to the rest of the stations.

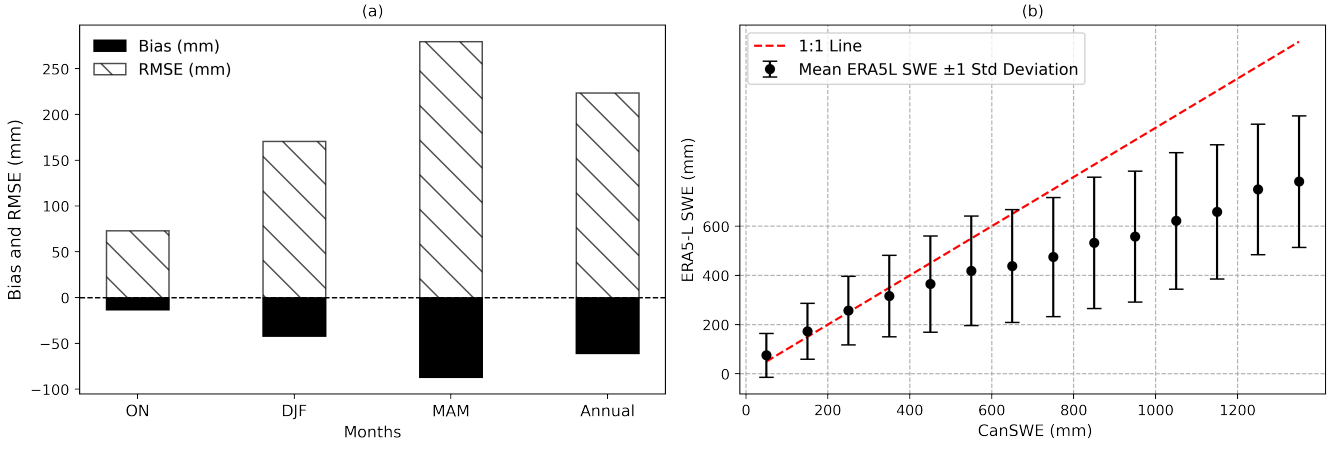

Figure 2. a) Root Mean Square Error (RMSE) and Bias in ERA5L SWE estimates across distinct seasons: Fall (ON), Winter (DJF), Spring (MAM), and the Entire Year. b) Comparative analysis of ERA5L against CanSWE, considering ± 1 standard deviation.

3.2 Performance of bias correction methods

MBS successfully reduced mean bias to zero when averaged over all time steps, yet seasonal biases persisted (Fig 4). The subtraction of negative bias inadvertently introduced positive biases, affecting seasons and locations characterized by low

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

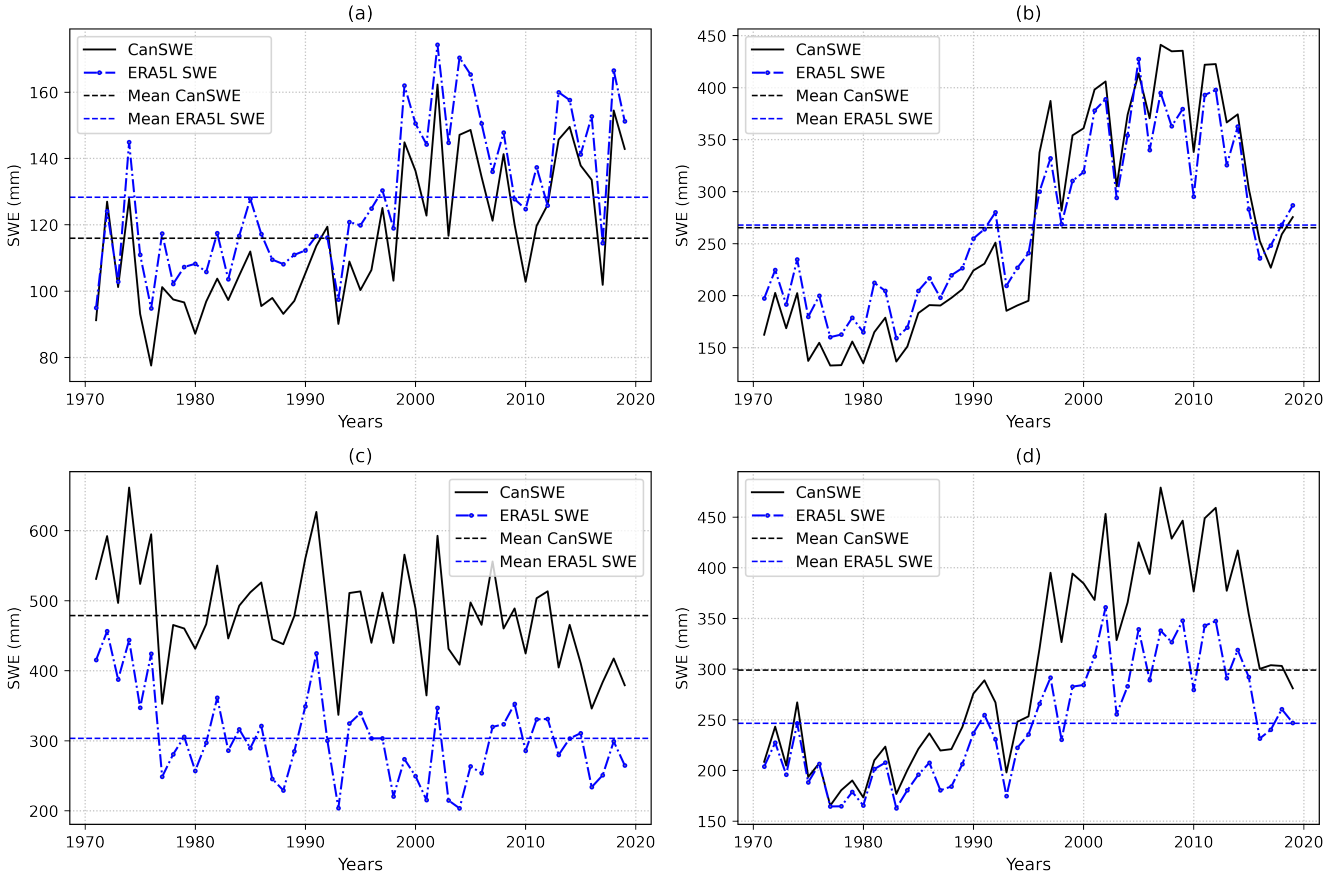

Figure 3. Interannual variability of Snow Water Equivalent (SWE) as observed by CanSWE and ERA5L across elevation categories: (a) Low elevation (<500m), (b) Mid elevation (500-1500m), and (c) High elevation (>1500m), along with (d) the entire study area. The dotted lines indicate the mean (1971-2019) SWE values.

SWE values. For instance, there is a noticeable positive bias in SWE during the Fall and early Winter months, as illustrated in Figure 5(b). These positive biases extend to annual SWE values at low-elevation sites, as depicted in Figure 6 (b). Overall, the annual RMSE over the study area was reduced by only 3.74% by using MBS.

Similarly, SLR exhibited a performance comparable to MBS, introducing positive biases during the Fall and early Winter months, as evident in Figure 4(a) and Figure 5 (b). While there was some improvement in spring biases, the overall reduction in annual RMSE was limited to 4.11%, as shown in Figure 4 (b).

In contrast, MLR benefitted from the addition of more predictors, leading to a significant reduction in biases and RMSE across all seasons compared to SLR (Figure 4). MLR introduced a positive bias for Fall months, as depicted in Figure 4(a) and Figure 5(b), but it significantly improved SWE estimates for Winter and Spring months, resulting in a substantial 16.36%

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

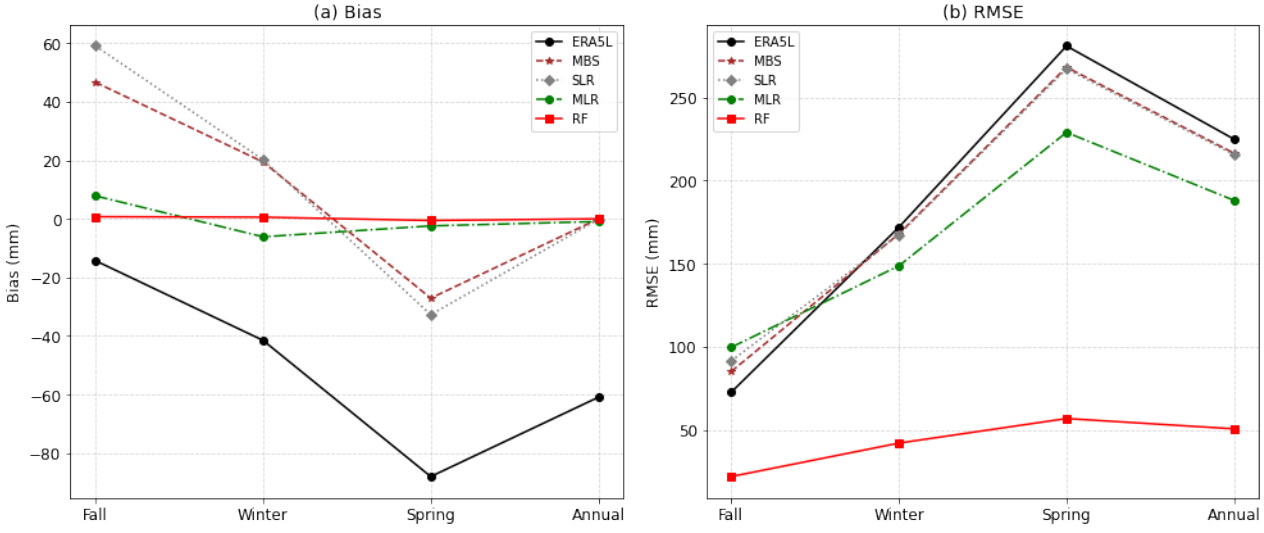

Figure 4. (a) Bias and (b) Root Mean Square Error (RMSE) comparison of ERA5L, MBS, SLR, MLR, and RF across different seasons (Fall, Winter, Spring, Entire Year).

reduction in annual RMSE. The scatterplot in Figure 7(a) reveals that the majority of data pairs indicate SWE values within the 0-250mm range across the study area. MLR captures approximately 42% of the variance in SWE (Figure 7a) but tends to overestimate SWE within the typical range (<250mm) and underestimate SWE values exceeding 500mm. The overestimation of low SWE values contributes to the positive bias introduced in the Fall months (Figure 4a). MLR's tendency to underestimate high SWE values is reflected in the underestimation of peak SWE (Fig 5b), the persistence of high dry bias in high-elevation sites (Fig 6c), and consistent underestimation of SWE values higher than 500mm at almost all elevation zones (Fig 8). However, MBS, SLR, and MLR demonstrated the ability to capture the inter-annual variability of SWE, albeit with biases, as illustrated in Figure 5(a).

The Random Forest (RF) model consistently demonstrates superior performance across training, validation, and test datasets. It demonstrated the lowest Root Mean Square Error (RMSE) and bias values across diverse seasons, affirming its robustness and reliability in capturing intricate patterns within the data. Remarkably, it achieved a 77.5% reduction in annual RMSE without introducing significant positive biases during the fall and winter months, as highlighted in Fig 4a and Fig 5b. Spatially, the sites with strong positive (>100mm) and negative (<-100mm) bias reduced from approximately 89 and 87 sites to 4 and 4 sites respectively. The Relative mean bias (RMB) reduced from 41% to 23%, with sites with RMB more than 100% reducing from 129 to 42 sites only. The number of sites with negative RMB reduced from 291 sites to 130 only. The scatterplot in Figure 7(b) reveals that RF captures approximately 97% of the variance in SWE, typically performing well in predicting SWE values 0-1000mm resulting in good performance at mid and high-elevation areas (Fig 8 b-d). Notably, the accuracy of RF predictions declines for very high SWE values (>1000mm) as depicted in Fig 7. This can be attributed to the lesser representation of the SWE values of more than 1000mm in the training data (4.6%).

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

The enhanced performance of RF can be attributed to its capability to capture the non-linearities between SWE and predictor variables. Figure 10 illustrates that both Multiple Linear Regression (MLR) and RF had similar performance (RMSE and R^2) when using only one predictor. However, the addition of elevation significantly improved RF predictions compared to MLR. This improvement stems from the non-linear relationship between SWE and elevation (Fig A4). Given that SWE distribution is influenced by elevation, seasonality, and longitude, these three features were identified as the top three secondary predictors in terms of importance (Figure 9). Other local attributes, such as Slope, Aspect, NDVI, and year, were ranked as the least influential features. It is essential to note that lower ranking does not necessarily equate to lower contribution, as feature importance methods are biased in favor of variables with many possible split points(Breiman 2001; Wright et al. 2017). The reason is that—just by chance— a variable with a larger number of potential split points naturally increases the chances of encountering a split that significantly reduces impurity. This occurrence remains true regardless of whether the variables are associated with the outcome or not.Li et al. (2016) caution against relying solely on feature importance for variable selection in predictive models

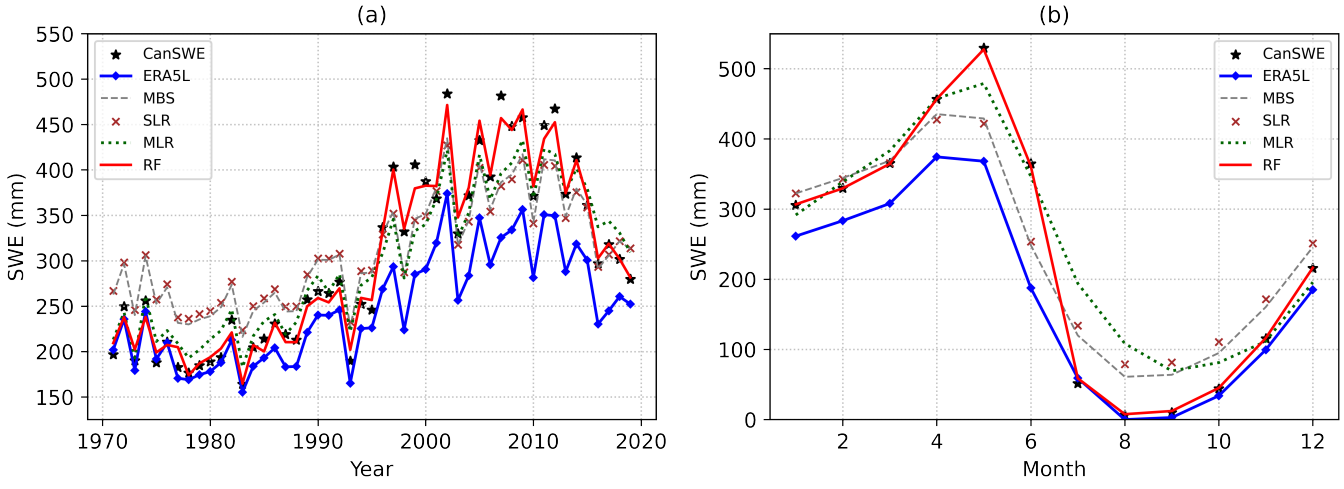

Figure 5. (a) Interannual and (b) Monthly variability of SWE estimates by CanSWE, ERA5L, MBS, SLR, MLR, and RF

RF has outperformed other techniques consistently while using random CV. However, the optimistic results obtained from RF in random CV indicate overfitting due to inherent spatial and temporal auto-correlation within the data. To quantify the impact of temporal overfitting, we trained the model on data spanning 1971-2014 and tested its predictive capabilities for the years 2015-2019 using leave-time-out cross-validation (LTOCV), the results of which are outlined in Table 3. While RF continues to outperform Multiple Linear Regression (MLR), its efficacy reduces with higher RMSE and reduced R2 compared to random CV. Since the years that went into training could only be predicted reliably (indicated by results on random CV), temporal overfitting can be assumed to be contributing to the random CV errors. To investigate the spatial overfitting, we employed leave-one-out cross-validation (LOOCV), iteratively training models on randomly selected 90% of sites and evaluating performance on the remaining 10%. Despite RF maintaining its superiority over MLR, its performance dips to 0.72, indicating

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

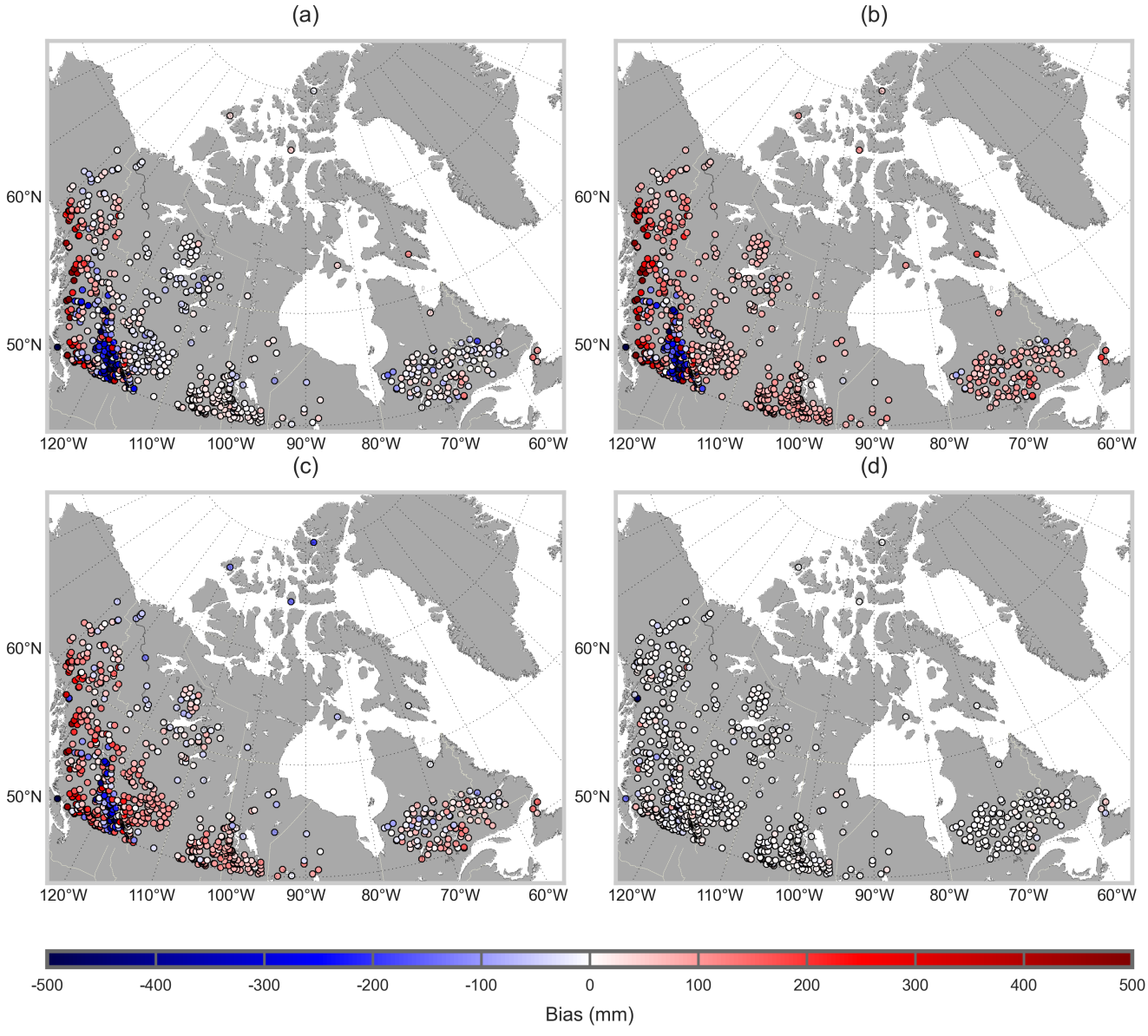

Figure 6. Spatial distribution of SWE biases (mm) in (a) ERA5L (b) ERA5L after MBS (c)ERA5L after bias correction by MLR (c) ERA5L after bias correction by RF

a more pronounced impact of spatial overfitting. It is not entirely surprising that MLR showed poorer performance in LTOCV compared to LOOCV, whereas RF exhibited better performance in LTOCV. This discrepancy likely arises from MLR's limitation in capturing temporal dependency, particularly when training and testing periods differ. This mismatch diminished the

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

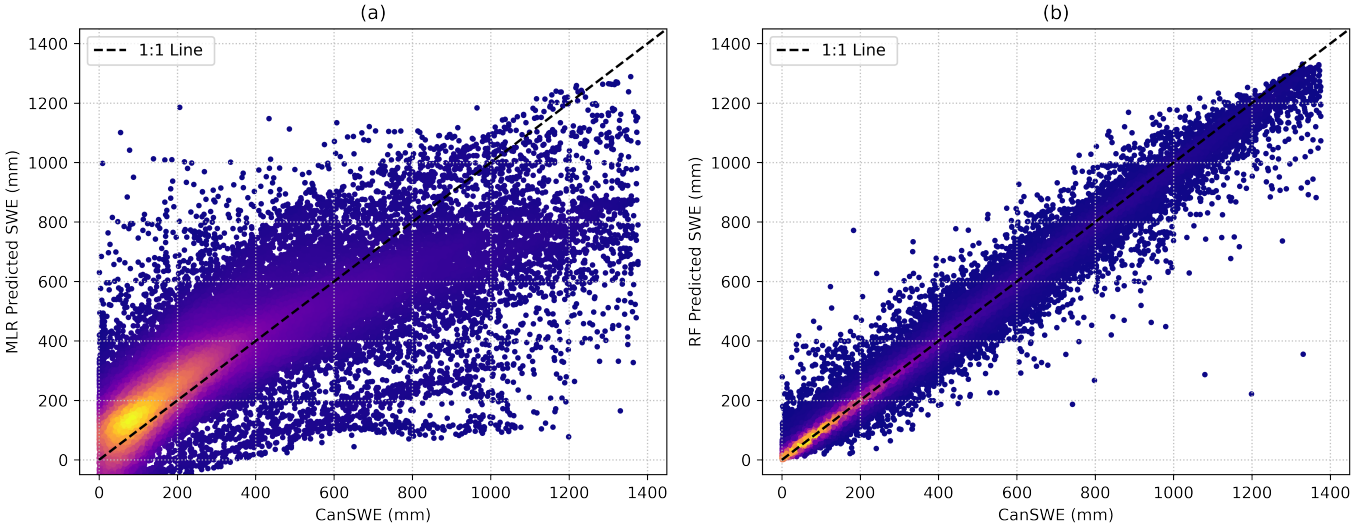

Figure 7. Scatter plot of CanSWE vs predicted SWE by (a) MLR and (b) RF. The dashed line represents y=x line

model performance as the relationships learned during training may not effectively extend to unseen time periods with altered correlations.

3.3 Spatial Transferability of the RF model

RF can be used as a tool to predict SWE for remote areas where SWE measurements are sparse. While Random CV offers insights into interpolation errors, our objective extends to evaluating extrapolation errors, specifically how well the model predicts on unsampled sites. We assessed the transferability of RF at distinct ecozones—defined as areas characterized by unique climates, vegetation, and ecosystems(Wiken 1986). Given the varying snow distribution across ecozones, the results revealed diverse performance across different zones (Fig 11).

In general, RF demonstrated proficiency in capturing the typical SWE range across most ecozones. However, it exhibited limitations in accurately predicting extremely low or high SWE values, particularly in ecozones such as the Arctic (characterized by extremely low SWE) and the Montane Cordillera and Pacific Maritime regions (both characterized by very high SWE). In particular, within the Montane Cordillera and Pacific Maritime ecozones, where elevation ranges from 200 to 2500m, Random Forest struggled to accurately predict high SWE values. As seen in Fig 11, the scatterplot has two distinct branches: one representing accurate predictions, typically observed in regions with moderate elevations within these zones, and the other indicating underprediction of high SWE values from very high elevations. This challenge may stem from the inadequate representation of extreme SWE values especially at very high elevations in the training dataset.

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

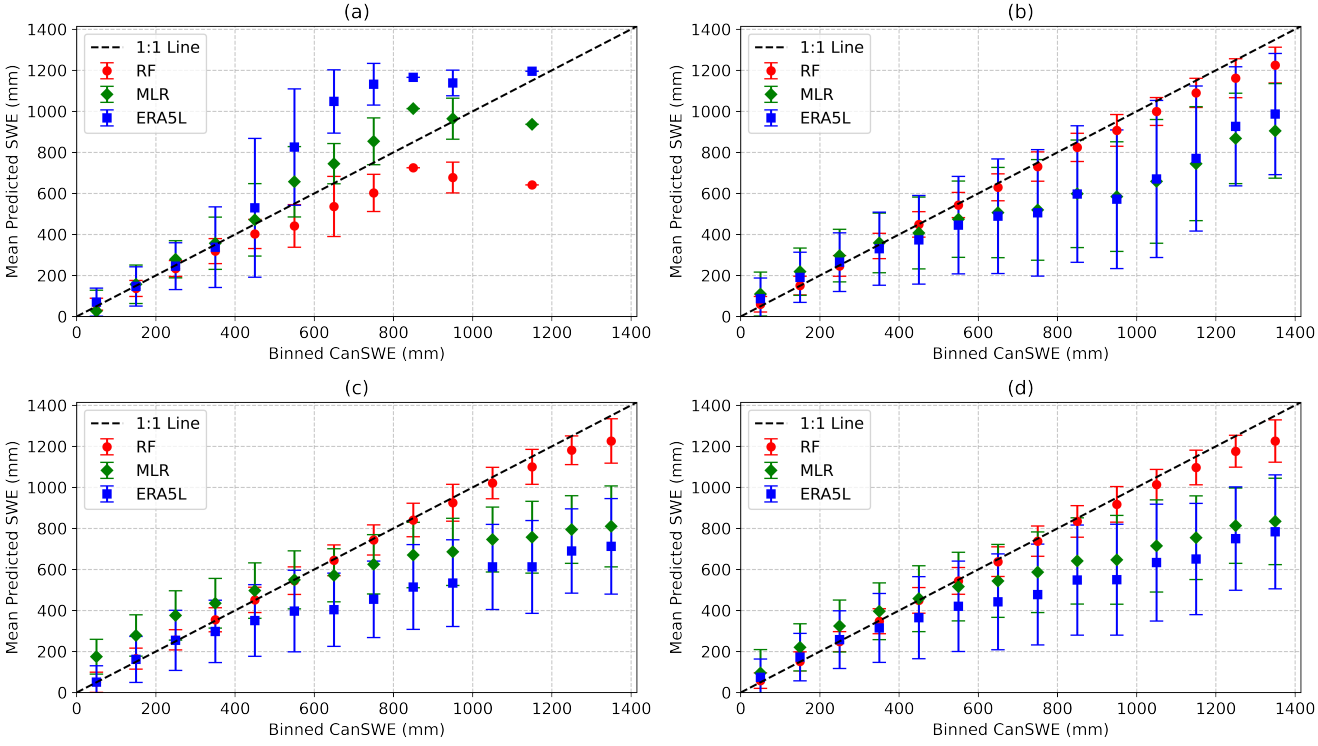

Figure 8. Performance of ERA5L, MLR, and RF predicted SWE versus CanSWE±1 at a) Low elevation (<500m) areas(b)Mid elevation (500-1500)areas, and (c) at High elevation areas (>1500m) (d) Entire Study Area.

Table 2. Comparison of model performance (RMSE and R^2) for linear regression and RF (LOOCV and LTOCV)

Validation Strategy	Metric	ERA5L	MLR	RF
LTOCV	RMSE	218.77	194.12	103.12
	R^2		0.34	0.87
LOOCV	RMSE	224.01	174.09	153.58
	R^2		0.68	0.74

4 Discussion

We discovered a negative bias in the ERA5L dataset, particularly in areas with high elevation (>1500m). Reanalysis datasets are susceptible to SWE biases due to limitations related to model physics and resolution(Snauffer et al. 2016). However, our finding of negative bias in very high elevations is in contrast to the findings of Kouki et al. (2023), who observed ERA5L to overestimate SWE in mountainous regions. The difference in results can be attributed to the use of different reference datasets. Kouki et al. (2023) utilized MERRA-2, Brown, and Crocus v7 SWE products as reference data. However, these datasets were found to underestimate SWE in Canada compared to reference snow survey datasets(Mortimer et al. 2020). In our study, we

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

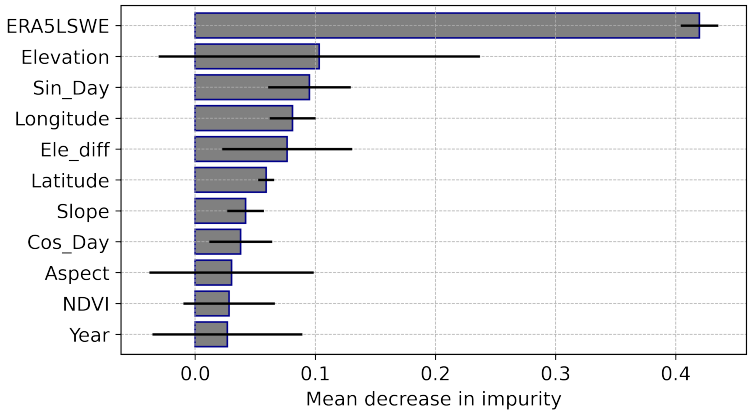

Figure 9. Feature importance results for the Random Forest model, with ±1 standard deviation indicated by the black error bars.

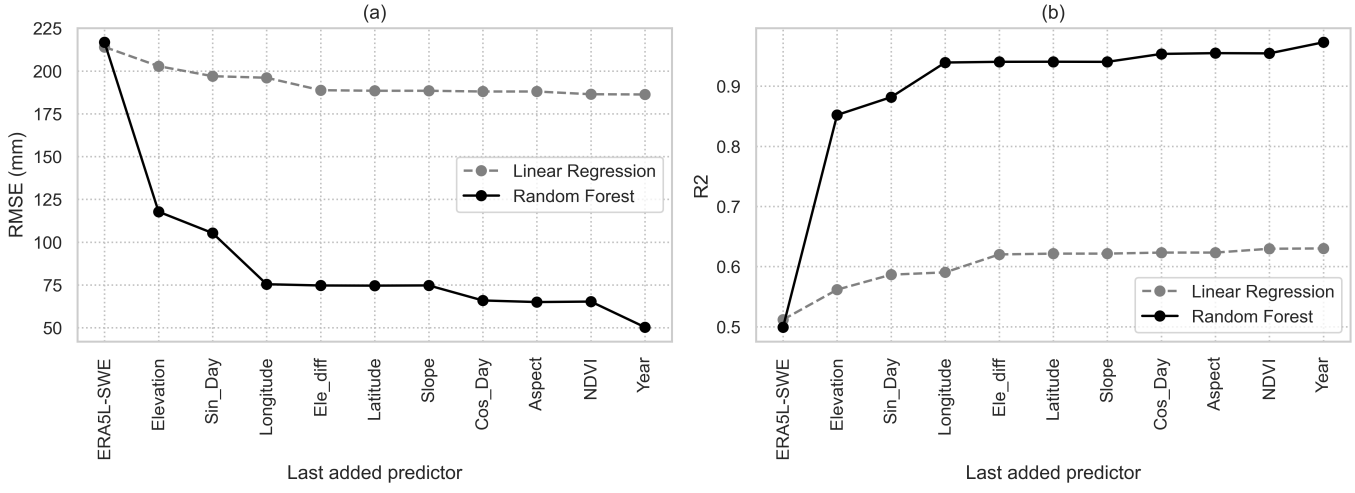

Figure 10. Improvements in (a)RMSE and (b) R^2 in MLR and RF model upon addition of predictors

relied on CanSWE data, which integrates in-situ SWE measurements from various sources such as snow surveys, snow pillows, and gamma radiation sensors(Vionnet et al. 2021), as a reference. While acknowledging the inherent uncertainties associated with these measurements—for instance, snow pillows may exhibit uncertainties ranging from 6-12% and gamma sensors may have a measurement error of approximately ±20mm(Taheri and Mohammadian 2022), we considered them as the ground truth for our study.

While MBS effectively zeroed the annual bias, it poorly addressed the seasonal SWE errors, consistent with the observations by King et al. (2020). The seasonal errors also persisted upon the application of linear regression. We also observed that linear techniques (SLR, MLR) predicted negative SWE values which is not possible in the physical world. King et al. (2020)

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

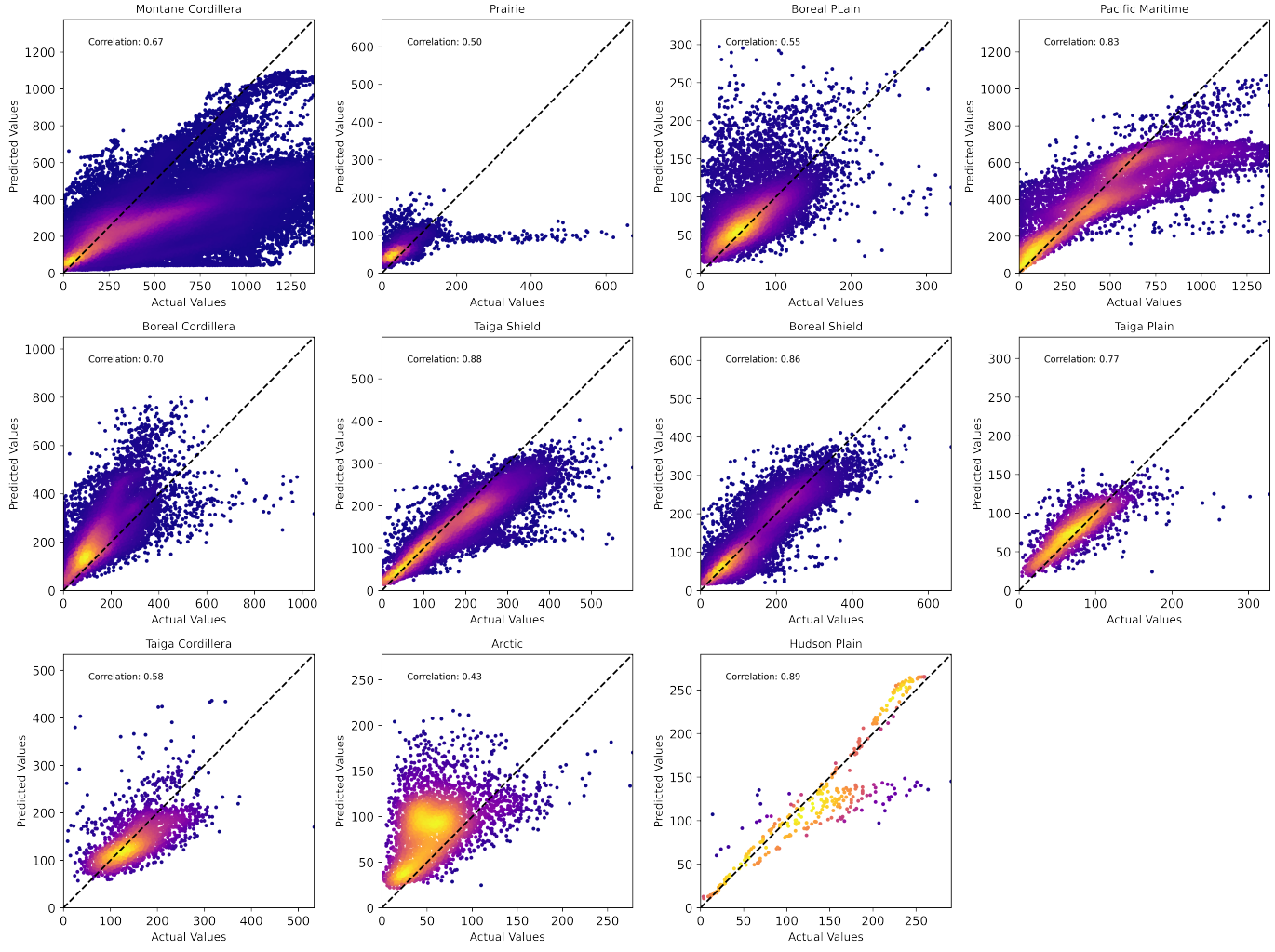

Figure 11. Scatter plot of RF Predicted vs CanSWE at different ecozones of Canada

also reported something similar about MBS and highlighted that bias correction techniques can alter the relationship between predictor and target variables, occasionally generating values that lack physical meaning.

RF consistently outperformed all methods, reducing the total RMSE by 77.5%. This aligns with previous studies, such as that byKing et al. (2020) and Bennett et al. (2022), affirming Random Forest as a robust and effective bias correction technique. However, despite an overall good performance, the Random Forest exhibited suboptimal performance on some coastal and high mountain sites (depicted by deeper colors in Fig 6d). The reasons for this behavior could be explored in future work.

As per our study, elevation was found to be the most important predictor after ERA5LSWE. This is consistent with the findings of Bennett et al. (2022) who found higher SWE accumulations at higher elevations. Interestingly, local site characteristics

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

like Slope, aspect, and NDVI were found to be the least important factors as per MDI. However, the accuracy of both models (MLR and RF) shows a non-monotonous increase with the addition of predictors (Fig 10). Moreover, the DEM we used for the calculation of site characteristics had a spatial resolution of 90m which might not be sufficient to capture the local microtopography. Finer scale DEM can be used in the future to investigate the importance of local features. Further, the overly optimistic results from random cross-validation suggest potential overfitting due to temporal and spatial autocorrelation. To address this, we employed blocking techniques to assess overfitting caused by correlated temporal and spatial data. Similar to the findings of Meyer et al. (2018), we found that the RF predictions were not reliable for the years and locations that were not included in the training data indicating impacts of spatial and temporal overfitting.

In terms of spatial transferability, RF did not perform well in ecoregions characterized by extremely low or high SWE values This highlights the struggle of RF to predict on a range outside of the training data despite its general ability to capture typical SWE values across most ecoregions. The observed limited transferability of the best model aligns with the conclusions drawn byMeyer et al. (2018) and Bjerre et al. (2022). This limitation is attributed to spatial block techniques that restrict the data available for model training, leading to excessively pessimistic estimates of the error. Since we aim to predict for both observed and unseen locations using the best model, these results offer a robust approximation of both interpolation and extrapolation errors.

5 Conclusions

We evaluated the performance of ERA5LSWE and bias correction methods in Northern Canada (land areas north of 50 °N). The evaluation was done against CanSWE data for the period 1971-2019. We found that ERA5L underestimates annual mean SWE averaged over the entire Northern Canada domain by approximately 61 mm but most underestimation was observed on high elevation (>1500m) sites. However, ERA5L was found to be overestimating SWE at low elevation sites (<500m). To correct these biases, we applied four correction methods, i.e. MBS, SLR, MLR, and RF. RF exhibited the most promising performance, reducing the RMSE by an impressive 77.5% and minimizing the annual mean bias from 61.2 mm to 0.01 mm. However, the accuracy of RF predictions declined for SWE values greater than 1000mm. We also found that RF results were impacted by both spatial and temporal overfitting as the accuracy of the RF model reduced when it predicted SWE for unseen sites and years. However, the impact of spatial overfitting was more pronounced. RF showed limited spatial transferability to regions characterized by very low or very high SWE values. However, it was able to capture the typical SWE values at most ecoregions. These findings underline the need for advanced bias correction techniques, addressing overfitting concerns, and enhancing spatial transferability across diverse geographical and climatic settings, particularly in regions with extreme snowpack conditions. Additionally, the impact of improved DEM resolution on the model performance needs to be explored in the future.

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

6 Acknowledgements

We thank Homa Kheyrollah Pour and Andre Erler for their valuable feedback, which significantly enhanced the quality of this paper. Additionally, we thank Fraser King, whose assistance with a portion of the code greatly facilitated our research.

Competing interests. The authors declare that they have no competing conflicts

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

References

Barnett, T. P., Adam, J. C., and Lettenmaier, D. P.: Potential impacts of a warming climate on water availability in snow-dominated regions, Nature, 438, 303–309, https://doi.org/10.1038/nature04141, number: 7066 Publisher: Nature Publishing Group, 2005.

Bennett, K. E., Miller, G., Busey, R., Chen, M., Lathrop, E. R., Dann, J. B., Nutt, M., Crumley, R., Dillard, S. L., Dafflon, B., Kumar, J., Bolton, W. R., Wilson, C. J., Iversen, C. M., and Wullschleger, S. D.: Spatial patterns of snow distribution in the sub-Arctic, The Cryosphere, 16, 3269–3293, https://doi.org/10.5194/tc-16-3269-2022, publisher: Copernicus GmbH, 2022.

Bjerre, E., Fienen, M. N., Schneider, R., Koch, J., and Højberg, A. L.: Assessing spatial transferability of a random forest metamodel for predicting drainage fraction, Journal of Hydrology, 612, 128 177, https://doi.org/10.1016/j.jhydrol.2022.128177, 2022.

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

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

Chen, Y., Sharma, S., Zhou, X., Yang, K., Li, X., Niu, X., Hu, X., and Khadka, N.: Spatial performance of multiple reanalysis precipitation datasets on the southern slope of central Himalaya, Atmospheric Research, 250, 105 365, https://doi.org/10.1016/j.atmosres.2020.105365, 2021.

Didan, K., Munoz, A. B., Solano, R., and Huete, A.: MODIS vegetation index user's guide (MOD13 series), 2015.

European Space Agency and Sinergise: Copernicus Global Digital Elevation Model, https://doi.org/10.5069/G9028PQB, accessed: 2024-02-21, 2021.

Fierz, C., Armstrong, R. L., Durand, Y., Etchevers, P., Greene, E., McClung, D. M., Nishimura, K., Satyawali, P. K., and Sokratov, S. A.: The International Classification for Seasonal Snow on the Ground, IHP-VII Technical Documents in Hydrology 83, UNESCO-IHP, Paris, France, http://www.unesco.org/water/ihp, iACS Contribution No. 1, 2009.

Hastings, D. A., Dunbar, P. K., Elphingstone, G. M., Bootz, M., Murakami, H., Maruyama, H., Masaharu, H., Holland, P., Payne, J., Bryant, N. A., Logan, T. L., Muller, J.-P., Schreier, G., and MacDonald, J. S.: The Global Land One-Kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0, 325 Broadway, Boulder, Colorado 80305-3328, U.S.A., 1999.

Imaoka, K., Kachi, M., Fujii, H., Murakami, H., Hori, M., Ono, A., Igarashi, T., Nakagawa, K., Oki, T., Honda, Y., and Shimoda, H.: Global Change Observation Mission (GCOM) for Monitoring Carbon, Water Cycles, and Climate Change, Proceedings of the IEEE, 98, 717–734, https://doi.org/10.1109/JPROC.2009.2036869, 2010.

Ishwaran, H.: The effect of splitting on random forests, Machine Learning, 99, 75–118, https://doi.org/10.1007/s10994-014-5451-2, 2015.

Kelly, R.: The AMSR-E Snow Depth Algorithm: Description and Initial Results, Journal of The Remote Sensing Society of Japan, 29, 307–317, https://doi.org/10.11440/rssj.29.307, 2009.

King, F., Erler, A. R., Frey, S. K., and Fletcher, C. G.: Application of Machine Learning Techniques for Regional Bias Correction of Snow Water Equivalent Estimates in Ontario, Canada, Hydrology and Earth System Sciences, 24, 4887–4902, 2020.

Kouki, K., Luojus, K., and Riihelä, A.: Evaluation of snow cover properties in ERA5 and ERA5-Land with several satellite-based datasets in the Northern Hemisphere in spring 1982-2018, The Cryosphere Discussions, pp. 1–33, https://doi.org/10.5194/tc-2023-53, publisher: Copernicus GmbH, 2023.

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

Li, J., Tran, M., and Siwabessy, J.: Selecting Optimal Random Forest Predictive Models: A Case Study on Predicting the Spatial Distribution of Seabed Hardness, PLOS ONE, 11, e0149 089, https://doi.org/10.1371/journal.pone.0149089, publisher: Public Library of Science, 2016.

Liaw, A. and Wiener, M.: Classification and regression by randomForest, R news, 2, 18–22, 2002.

Meromy, L., Molotch, N. P., Link, T. E., Fassnacht, S. R., and Rice, R.: Subgrid variability of snow water equivalent at operational snow stations in the western USA, Hydrological Processes, 27, 2383–2400, https://doi.org/10.1002/hyp.9355, 2012.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T.: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation, Environmental Modelling & Software, 101, 1–9, https://doi.org/10.1016/j.envsoft.2017.12.001, 2018.

Minola, L., Zhang, F., Azorin-Molina, C., Pirooz, A. A. S., Flay, R. G. J., Hersbach, H., and Chen, D.: Near-surface mean and gust wind speeds in ERA5 across Sweden: towards an improved gust parametrization, Climate Dynamics, 55, 887–907, https://doi.org/10.1007/s00382-020-05302-6, 2020.

Mortimer, C., Mudryk, L., Derksen, C., Luojus, K., Brown, R., Kelly, R., and Tedesco, M.: Evaluation of long-term Northern Hemisphere snow water equivalent products, The Cryosphere, 14, 1579–1594, https://doi.org/10.5194/tc-14-1579-2020, publisher: Copernicus GmbH, 2020.

Muñoz-Sabater, 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.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, publisher: Copernicus GmbH, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., and et al.: Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.02881, 2017.

Shao, D., Li, H., Wang, J., Hao, X., Che, T., and Ji, W.: Reconstruction of a daily gridded snow water equivalent product for the land region above 45°N based on a ridge regression machine learning approach, Earth System Science Data, 14, 795–809, https://doi.org/10.5194/essd-14-795-2022, publisher: Copernicus GmbH, 2022.

Snauffer, A. M., Hsieh, W. W., and Cannon, A. J.: Comparison of gridded snow water equivalent products with in situ measurements in British Columbia, Canada, Journal of Hydrology, 541, 714–726, https://doi.org/10.1016/j.jhydrol.2016.07.027, 2016.

Snauffer, A. M., Hsieh, W. W., Cannon, A. J., and Schnorbus, M. A.: 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, publisher: Copernicus GmbH, 2018.

Taheri, M. and Mohammadian, A.: An Overview of Snow Water Equivalent: Methods, Challenges, and Future Outlook, Sustainability, 14, 11 395, https://doi.org/10.3390/su141811395, number: 18 Publisher: Multidisciplinary Digital Publishing Institute, 2022.

Tedesco, M. and Jeyaratnam, J.: AMSR-E/AMSR2 Unified L3 Global Daily 25 km EASE-Grid Snow Water Equivalent, Version 1, Tech. rep., NASA National Snow and Ice Data Center Distributed Active Archive Center, Boulder, Colorado, USA, URL, 2019.

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

Velikou, K., Lazoglou, G., Tolika, K., and Anagnostopoulou, C.: Reliability of the ERA5 in Replicating Mean and Extreme Temperatures across Europe, Water, 14, 543, https://doi.org/10.3390/w14040543, number: 4 Publisher: Multidisciplinary Digital Publishing Institute, 2022.

Vionnet, V., Mortimer, C., Brady, M., Arnal, L., and Brown, R.: Canadian Historical Snow Water Equivalent Dataset (CanSWE, 1928–2020), Earth System Science Data, 13, 4603–4619, 2021.

Wiken, E. B.: Terrestrial Ecozones of Canada, Environment Canada, Lands Directorate, 1986.

Wright, M. N., Dankowski, T., and Ziegler, A.: Unbiased split variable selection for random survival forests using maximally selected rank statistics, Statistics in Medicine, 36, 1272–1284, 2017.

Zhao, P. and He, Z.: A First Evaluation of ERA5-Land Reanalysis Temperature Product Over the Chinese Qilian Mountains, Frontiers in Earth Science, 10, https://www.frontiersin.org/articles/10.3389/feart.2022.907730, 2022.

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

Appendix A

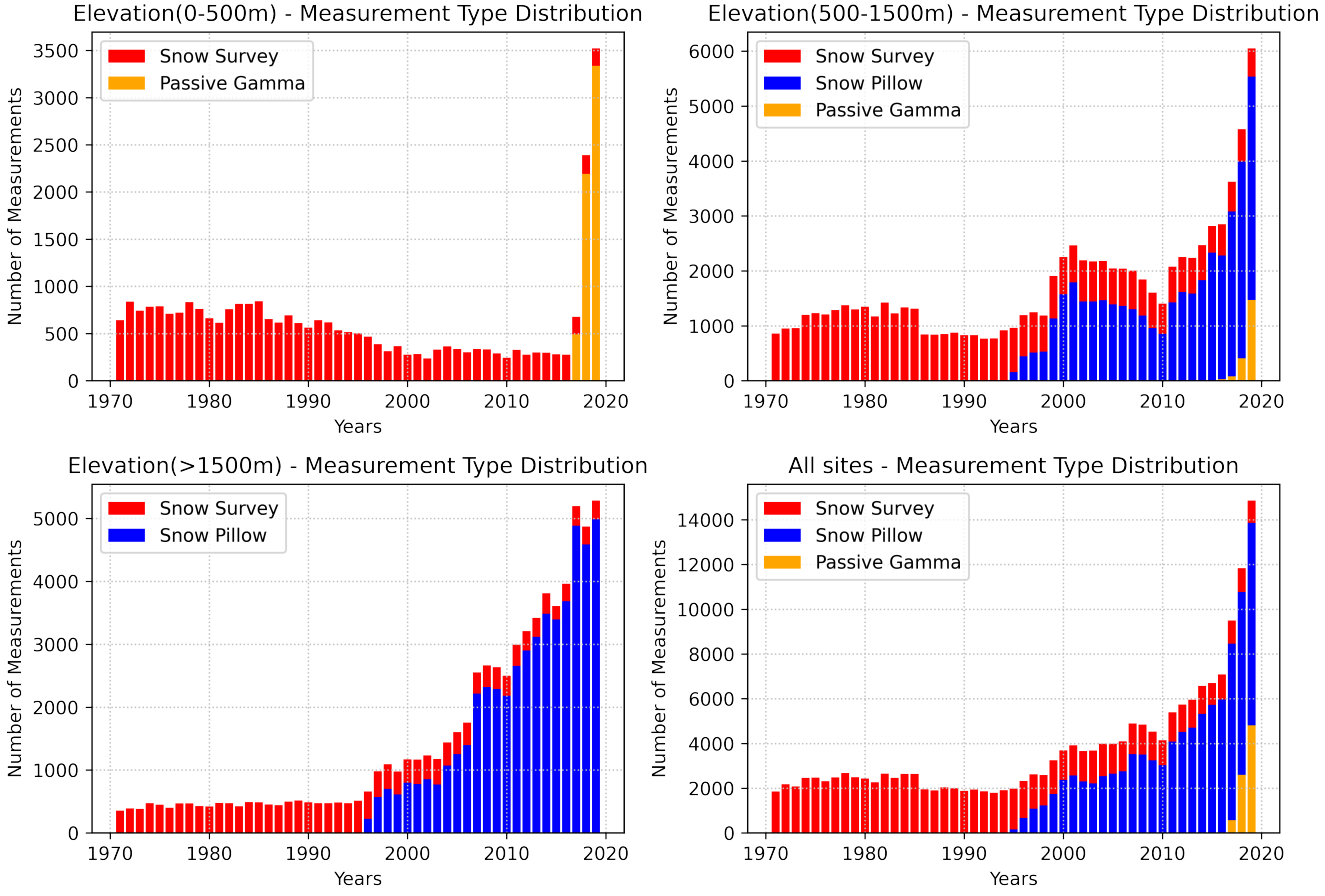

Figure A1. Distribution of types of SWE measurements,i.e. Snow Survey, Snow Pillows and Passive Gamma sensors at different elevation zones

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

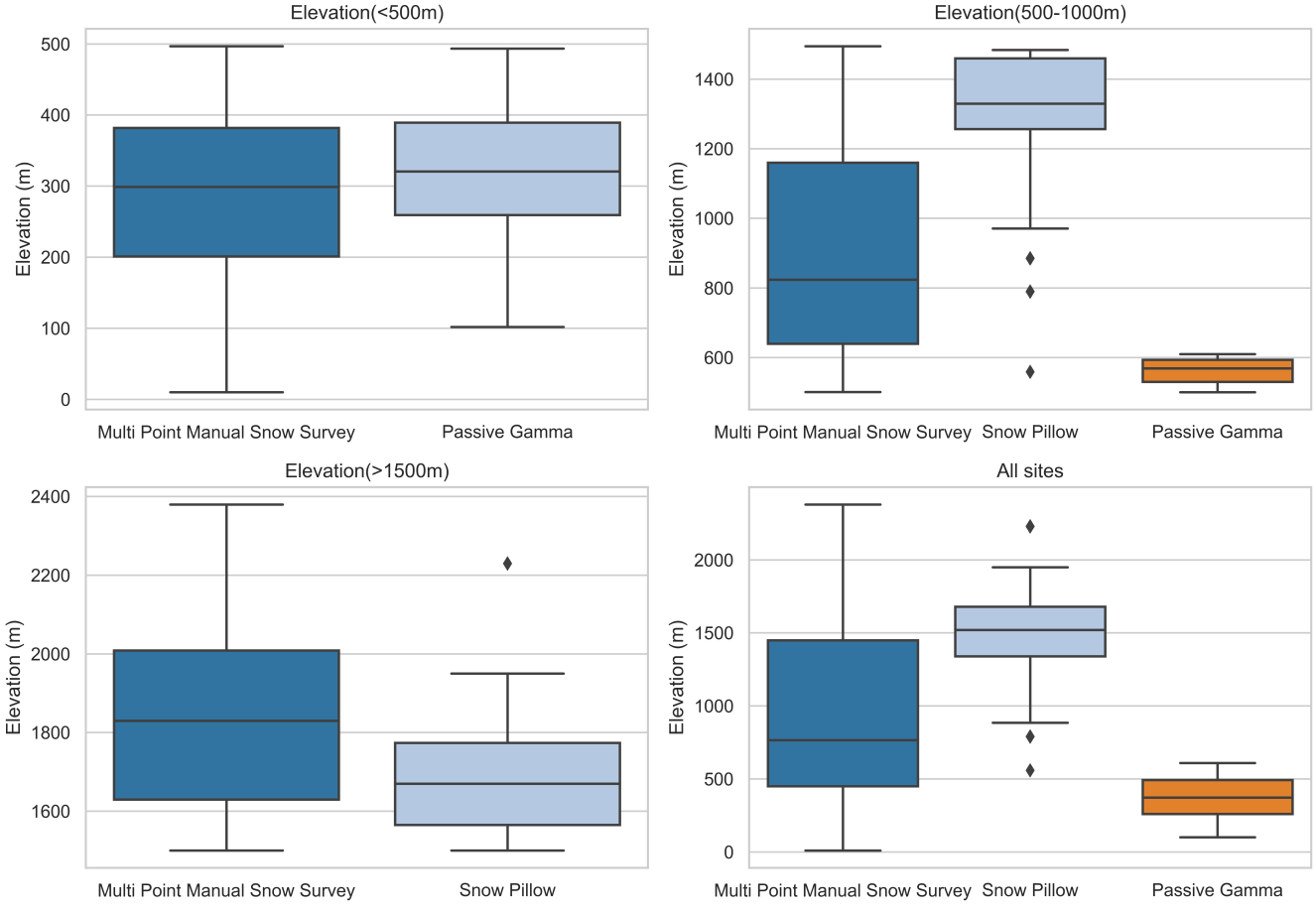

Figure A2. Distribution of elevation of SWE measurements at different elevations zones

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

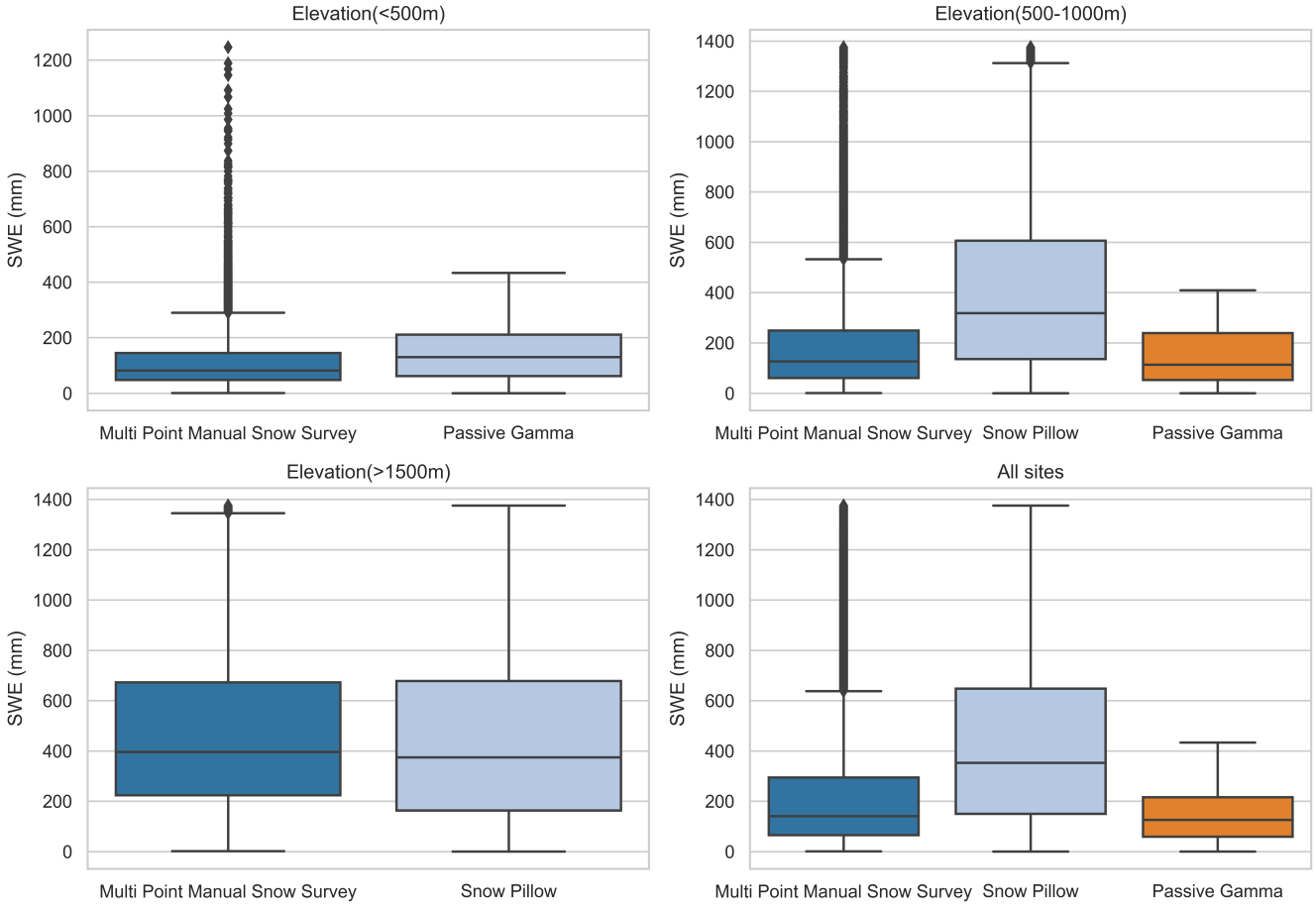

Figure A3. Distribution of SWE values by different measurement types at different elevations zones

https://doi.org/10.5194/egusphere-2024-639
Preprint. Discussion started: 27 March 2024

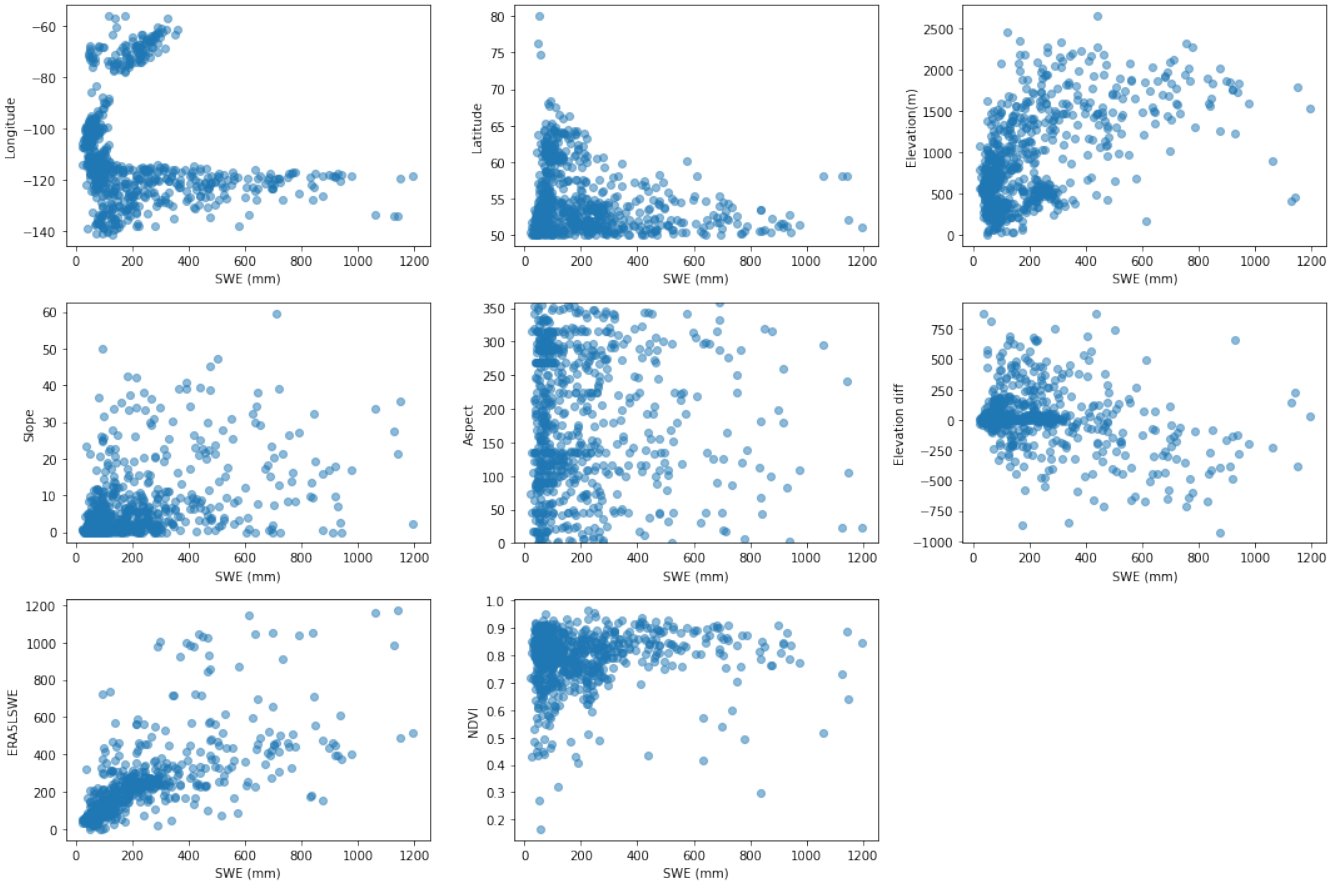

Figure A4. Scatter plot depicting relationship between target variable (SWE) and and Precdictor variables