8-7-2024

Bootstrap Methods for Bias-Correcting Probability Distribution Parameters Characterizing Extreme Snow Accumulations

Kenneth Pomeyie
Utah State University, kenneth.pomeyie@usu.edu

Brennan Bean
Utah State University, brennan.bean@usu.edu

Recommended Citation

Pomeyie, K.; Bean, B. Bootstrap Methods for Bias-Correcting Probability Distribution Parameters Characterizing Extreme Snow Accumulations. *Glacies* **2024**, *1*, 35-56. https://doi.org/10.3390/glacies1010004

glacies

MDPI

Article

Bootstrap Methods for Bias-Correcting Probability Distribution Parameters Characterizing Extreme Snow Accumulations

Kenneth Pomeyie * and Brennan Bean

Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT 84322, USA; brennan.bean@usu.edu

* Correspondence: kenneth.pomeyie@usu.edu

Abstract: Accurately quantifying the threat of collapse due to the weight of settled snow on the roof of a structure is crucial for ensuring structural safety. This quantification relies upon direct measurements of the snow water equivalent (SWE) of settled snow, though most weather stations in the United States only measure snow depth. The absence of direct load measurements necessitates the use of modeled estimates of SWE, which often results in the underestimation of the scale/variance parameter of the distribution of annual maximum SWE. This paper introduces a novel bias correction method that employs a bootstrap technique with regression-based models to calibrate the variance parameter of the distribution. The efficacy of this approach is demonstrated on real and simulated datasets. The findings reveal varied levels of success, with the efficacy of the proposed approach being inherently dependent on the quality of the selected regression-based model. These findings demonstrate that integrating our approach with a suitable regression-based model can produce unbiased or nearly unbiased annual maximum SWE distribution parameters in the absence of direct SWE measurements.

Keywords: bias correction; extreme value analysis; snow water equivalent; snow depth

Citation: Pomeyie, K.; Bean, B. Bootstrap Methods for Bias-Correcting Probability Distribution Parameters Characterizing Extreme Snow Accumulations. *Glacies* **2024**, *1*, 35–56. https://doi.org/10.3390/glacies1010004

Academic Editor: Steven R. Fassnacht

Received: 28 June 2024
Revised: 16 July 2024
Accepted: 17 July 2024
Published: 7 August 2024

1. Introduction

Accurately predicting and understanding environmental hazards is essential for building resilient infrastructure, particularly in regions subject to extreme weather. One hazard that demands significant attention in some parts of the Conterminous United States (CONUS) is the weight of settled snow on the roof of a structure. Quantifying the threat of settled snow to structural safety requires precise estimations of the probability distribution representing annual maximum snow load, or, equivalently, the snow water equivalent (SWE) [1]. However, direct observations of the SWE are currently limited across the CONUS.

One way to overcome the SWE data scarcity involves the use of climate reanalysis products. Climate reanalysis products synthesize observed data and modeled output into gridded maps of climate measurements across a region of interest. These products allow researchers to retroactively examine weather patterns over extended periods, and are especially helpful in geographical areas where direct observations of environmental variables are limited or non-existent. Despite their utility, climate reanalysis products come with certain shortcomings. One issue is the difficulty in quantifying the uncertainty inherent within reanalysis data, as pointed out by Dick et al. (2022) [2].

Another way to overcome data scarcity is to estimate the SWE from weather stations measuring only snow depth. The use of weather stations measuring snow depth greatly increases the number of available weather stations for use in future mapping procedures. The SWE estimation models leverage covariate information, such as snow depth, temperature, elevation, and other climate factors, to estimate the SWE at locations only recording snow depth (see [3–5] for examples). Despite their usefulness, one common limitation of

these SWE estimation models is the underestimation of the variance of the estimated SWE. This limitation is described and demonstrated in Figure 4 of Wheeler et al. (2022) [6].

The variance underestimation issue of modeled SWE is inevitable mathematically. To demonstrate this, let $\boldsymbol{Y}$ be an $n \times 1$ column vector that represents SWE and let $\boldsymbol{X}$ represent an $n \times p$ covariate matrix of other climate and snow information that could be used to estimate the SWE. We can formulate the regression model as follows:

$$\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon}$ is an $n \times 1$ column vector that is assumed to consist of independent and identically distributed (iid) realizations from a normal distribution with zero mean and constant variance (σ_ϵ^2). The function $f(\boldsymbol{X})$ represents the deterministic part of the model, which could be a linear or non-linear function of the covariates in $\boldsymbol{X}$. A standard regression analysis tries to estimate the conditional mean $E(\boldsymbol{Y}|\boldsymbol{X})$ by minimizing the residual sum of squares for the available data. By the law of total expectation, as presented in [7], the mean of $\boldsymbol{Y}$ is $E(\boldsymbol{Y}) = E(E(\boldsymbol{Y}|\boldsymbol{X}))$. On the other hand, the variance of $\boldsymbol{Y}$ can be represented as follows:

$$\begin{aligned} Var(\boldsymbol{Y}) &= Var(E(\boldsymbol{Y}|\boldsymbol{X})) + E(Var(\boldsymbol{Y}|\boldsymbol{X})) \\ &= Var\Big(f(\hat{\boldsymbol{X}})\Big) + E(\sigma_\epsilon^2) \\ &= Var\Big(f(\hat{\boldsymbol{X}})\Big) + \sigma_\epsilon^2. \end{aligned} \tag{2}$$

This formulation implies that the variance of the predicted SWE, i.e., $\hat{\boldsymbol{Y}} = f(\boldsymbol{X})$, will be

$$Var(\hat{\boldsymbol{Y}}) = Var\Big(f(\hat{\boldsymbol{X}})\Big) < Var(\boldsymbol{Y}). \tag{3}$$

In extreme value analysis, accurately capturing the scale and shape of the distribution is crucial for characterizing rare events. Relying solely on the predictions of the regression model, which systematically underestimates the variance of SWE, can lead to inadequate characterizations of the weight of snow in design. To demonstrate this challenge, we consider Table 1, which showcases the impact of varying the standard deviation (σ) of a log-normal distribution on the 98th and 99th percentiles. The table reveals that even small changes in the estimated variance of the distribution result in substantial differences in the estimates of extreme percentiles. This emphasizes the sensitivity of extreme event estimation to changes in estimated variance in the distribution of the annual maximum SWE.

Table 1. Examining the impact of standard deviation variations in a log-normal distribution on the 98th and 99th percentiles. The mean parameter (μ) of the distribution is fixed at zero, while $\sigma = 1$ is used as a reference point for comparison.

	98th Percentile		**99th Percentile**	
σ	**Extreme Event**	**Relative Increase (%) (from $\sigma = 1$)**	**Extreme Event**	**Relative Increase (%) (from $\sigma = 1$)**
1	7.8		10.24	
1.1	9.57	21.1	12.9	26
1.2	11.8	51.2	16.3	59.1
1.3	14.4	84.6	20.6	100.1

This paper addresses the underestimation of variance in SWE estimation models by introducing a residual bootstrap bias correction technique. We demonstrate in Section 3 that our approach is easier to generalize and scale as compared to existing approaches. This paper also compares station-level distribution estimates with and without our bootstrap correction. In addition to this, we compare the station-level distribution of climate reanal-

ysis data with and without the use of the Empirical Cumulative Distribution Function (ECDF) correction method, as previously introduced by Cho and Jacobs (2020) [8].

In plain language terms, predictions of the SWE using snow depth have less variability than true measurements of the SWE, which negatively impacts probability distribution estimates of extreme snow accumulation. To solve this issue, we propose a new approach that creates simulated alternatives of SWE predictions that re-introduce the variability of the SWE model residuals back into the distribution fitting process. By simulating enough of these alternatives, we are able to generate new estimates for the distribution parameters that better account for the variability lost in the SWE prediction process as compared to the original estimates.

This paper proceeds as follows: We first present the background literature on bias correction and reliability-targeted loads in Sections 2.1 and 2.2. We then detail the process of our proposed residual bootstrap technique for correcting bias in the distribution scale in Section 3 and demonstrate its effectiveness through a set of simulations and real-world data analysis in Sections 4 and 5. The paper concludes with a discussion of the implications of our findings, limitations, and potential areas of future research in Sections 5.4 and 6. Note that all analyses were performed in R 4.4.0 [9] with the help of the tidyverse suite of packages [10], as well as the caret [11], distfixer [12], extRemes [13], fitdistrplus [14], GOFKernel [15], maps [16], pacman [17], ranger [18], future [19], gbm [20], kernlab [21], raster [22], and sf [23] R packages.

2. Literature Background

2.1. Bias Correction

Bias denotes systematic overestimations or underestimations by a model, and is often the result of a model's simplifying assumptions. This underscores the fact that bias is often tied to the model itself rather than the data it utilizes. A prime example of such model-centric bias is evident in ensemble tree machine learning models. These models have an inherent tendency to overestimate smaller values and underestimate larger values, a phenomenon independent of the data fed into the model [24].

Bias correction can be approached at either the point scale or the distribution scale. The point-scale approach seeks to ensure that the expectation of individual estimates matches the expectation of the observed values (e.g., Figure 1). Several research efforts have been made to mitigate the effects of point-scale bias in machine learning models. For example, Hooker and Mentch (2016) [25] incorporated an approach based on residual bootstrap to reduce bias for ensemble methods. The proposed method led to both substantial reductions in bias and improvements in predictive accuracy. Similarly, Breiman (1999) [26] proposed an adaptive bagging technique to reduce prediction bias in ensemble models. Despite its strengths, this technique has the potential limitation of an increased variance that can offset the decrease in bias pointed out by Breiman (1999) [26]. Additionally, the method can be computationally expensive due to the iterative training of multiple models, which may also result in a final model that is difficult to interpret. Ghosal and Hooker (2020) [27] redefined the approach by using two random forest (RF) models in a method called one-step boosted forests (OSBF). In OSBF, a second RF model is trained on the residuals of the initial forest. The second model is then used to correct the bias in the original model. Several studies [24,27,28] have shown that OSBF outperforms traditional RF models. Nonetheless, when the ultimate objective is to mitigate bias in estimates of the second-order moments (i.e., variance or scale) of a distribution, these techniques may not yield satisfactory results, as pointed out by Belitz and Stackelberg (2021) [29].

On the other hand, the distribution-scale bias correction approach, also known as the CDF matching method (CDFMM), aims to ensure that the distribution of estimated values (model distribution) matches the distribution of observed values (observational distribution). The approach is accomplished using a transfer function ($t(x)$), where an estimated value (x_{est}) is first transformed into probabilities ($p = F_{est}(x_{est})$) using the CDF of model distribution. Next, the probabilities are back-transformed to a corrected observation

($x_{cor} = F_{obs}^{-1}(p)$) using the inverse CDF of the observational distribution, where p ϵ [0, 1]. Mathematically, the approach is expressed as follows:

$$x_{cor} = \mathrm{t}(x_{est}) = F_{obs}^{-1}(F_{est}(x_{est})). \quad (4)$$

An example of this approach is demonstrated in Cho and Jacobs (2020) [8], where a 4 km SWE reanalysis product developed by the University of Arizona [30], hereafter referred to as UA SWE, is bias corrected using the 1 km Snow Data Assimilation System (SNODAS) project [31] developed by the National Weather Service's (NWS) National Operational Hydrologic Remote Sensing Center (NOHRSC). The UA SWE data were created through the assimilation of numerous ground-based measurements of the SWE from the Snowpack Telemetry (SNOTEL) and the NWS Cooperative Observer Program (COOP) network sites and 4 km gridded PRISM precipitation and temperature data. The SNODAS product (hereafter, SNODAS SWE), on the other hand, aims to integrate snow data from satellites, airborne platforms, and ground stations with model estimates of snow cover [32].

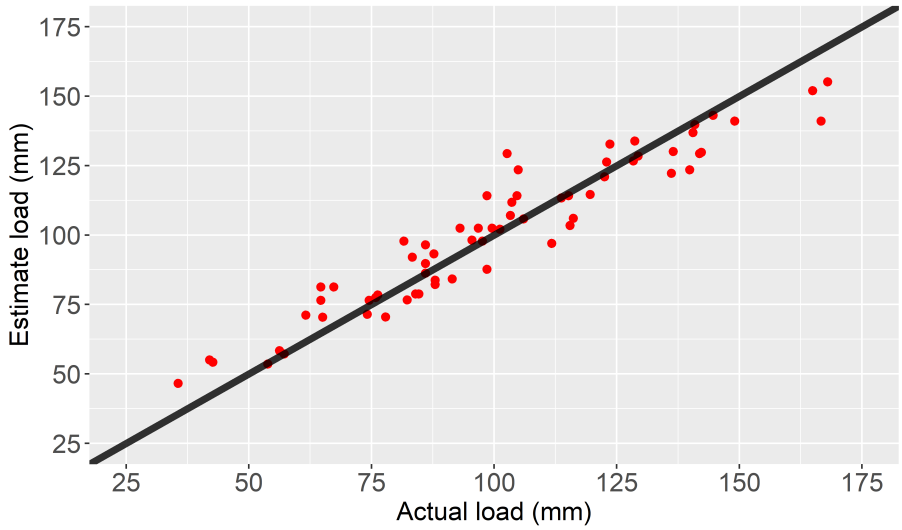

Figure 1. A scatter plot showcasing the systematic bias associated with the random forest model at the Grassy Lake weather station in Wyoming. The estimated load (in millimeters) represents the estimated snow load predicted using a random forest model, and the actual load (in millimeters) represents the observed ground snow load.

In the Cho and Jacobs (2020) [8] study, the bias correction process considers the annual maximum SNODAS SWE data (1 km $\times$ 1 km aggregated to 4 km $\times$ 4 km spatial grid) from October 2003 to May 2017 (14 snow years) and the annual maximum UA SWE data (4 km $\times$ 4 km spatial grid) from October 1981 to September 2017 (36 snow years). Although SNODAS provides reliable SWE data, its limited period of record restricts its use in extreme value analysis. On the other hand, UA SWE covers 36 years but is often regarded as less accurate. Therefore, the CDFMM is employed to bias correct UA SWE using SNODAS SWE data. At the grid cell level, the bias correction process involves transforming the UA SWE values using the UA and SNODAS SWE distributions developed for the 14 years where both datasets overlap. The transformation is applied as follows:

$$\mathrm{UA}_i^* = \mathrm{CDF}^{-}_{\mathrm{SNODAS}^{\mp}}(\mathrm{CDF}_{\mathrm{UA}^{\mp}}(\mathrm{UA}_i)), \text{ if } \mathrm{UA}_i \leq \max(\mathrm{UA}^{\mp}) \quad (5)$$

where UA_i^* and UA_i represent the bias-corrected and original UA SWE, respectively, at year i, $\mathrm{CDF}^{-}_{\mathrm{SNODAS}_{\mp}}$ and $\mathrm{CDF}_{\mathrm{UA}_{\mp}}$ represent the inverse CDF of SNODAS SWE and CDF of UA SWE for the overlapping years, and $\max(\mathrm{UA}_{\mp})$ is the maximum UA SWE in the overlapping years.

If UA_i at year i exceeds $\max(\mathrm{UA}_{\mp})$, the transformation is applied as follows:

$$\mathrm{UA}_i^* = max(\mathrm{SNODAS}_{\mp}) + [\mathrm{UA}_i - max(\mathrm{UA}_{\mp})] \times \frac{\sigma(\mathrm{SNODAS}_{\mp})}{\sigma(\mathrm{UA}_{\mp})} \quad (6)$$

where $\max(\mathrm{SNODAS}_{\mp})$ is the maximum SNODAS SWE in the overlapping years, and $\sigma(\mathrm{SNODAS}_{\mp})$ and $\sigma(\mathrm{UA}_{\mp})$ represent the standard deviation of SNODAS and UA SWE, respectively, in the overlapping years.

Although the mapping approach described in the above equations utilizes an empirical process to construct the transfer function, it is important to note that transfer functions can also be constructed using a parametric approach. In the empirical approach, the Cumulative Distribution Function (CDF) is defined by creating a quantile–quantile (Q-Q) plot which compares observed and modeled data. A transfer function is then generated through linear interpolation between the points on the plot. This approach, commonly known as Empirical CDF (ECDF) mapping, has been widely employed in hydrological studies to correct bias in climate model outputs [8,33,34]. Alternatively, the quantile mapping (QM) method has been used to define the CDF by estimating a set of quantiles for the observed and modeled data. A transfer function can then be created by interpolating between the corresponding quantile values [35]. Furthermore, the probability mapping method has been extensively used in climate modeling studies to correct bias [36–38]. This method involves fitting a parametric distribution to the observed and modeled data. The resulting transfer function is then formed by combining the estimated analytic CDF and the quantile function derived from the parametric distribution. See Lakshmanan et al. (2015) [39] for a comprehensive summary of distribution mapping techniques, including both empirical and parametric approaches.

However, all these mapping techniques require access to both the estimated and observed data. This means that weather stations that do not have co-located records of snow depth and SWE cannot be bias corrected for use in extreme value analysis using the ECDF approach. For this reason, we propose a distribution-scale bias correction approach based on empirical models using residual bootstrap. Our approach allows bias correction for cases where observed data are not available, as will be demonstrated in the following sections.

2.2. Reliability-Targeted Loads

Before the introduction of ASCE 7-22 (2022) [40] by the American Society of Civil Engineers (ASCE), ground snow loads used for designing buildings were based on a uniform probability of exceedance, i.e., a 50-year mean recurrence interval (MRI). Traditionally, the 50-year MRI snow load has been linked to structural reliability targets with calibrated safety factors. The pioneering calibration effort was carried out by Ellingwood et al. (1980) [41], with more details given by Ellingwood et al. (1982) [42] and Galambos et al. (1982) [43]. Although the initial calibration was correct, research by DeBock et al. (2017) [44] and Liel et al. (2017) [45] reveals how variations in regional snow accumulation patterns can result in site-specific snow loads either surpassing or falling short of reliability targets. This has led to a shift in recent engineering codes, moving from the MRI-based (uniform hazard) method towards "reliability-targeted" design snow loads, which target a uniform risk of structure failure.

The objective of the reliability-targeted load is to simultaneously consider the uncertainty of the environmental hazard and the uncertainty of the structural resistance to the hazard. This is accomplished via Monte Carlo simulation, where failure is defined as any instance where the simulated demand exceeds the simulated structural capacity. The design snow load is the value that ensures that the simulated risk of failure is at or below the failure thresholds defined in Chapter 1 of ASCE 7-22 (2022) [40]. The simulation process is iterated for various design ground snow loads to determine the achieved reliability index for a specific location of interest. Figure 2.10 of Bean et al. (2021) [46] shows a workflow of the reliability-targeted load simulation process.

The model for the reliability-targeted snow loading is derived from the design equation as follows:

$$1.2D + 1.0S = \phi R \tag{7}$$

where D represents dead load, S represents roof snow load, and ϕR represents structural resistance. These three variables are each characterized as random variables with associated probability distributions. Of interest in this paper is the distribution of roof snow load, which is modeled as a product of two random variables:

$$S = G_l \cdot G_r \tag{8}$$

Here, G_l is the random variable describing the ground snow load, and G_r describes the ground-to-roof conversion factor. The distribution of G_r is described in detail in Bean et al. (2021) [46]. In ASCE 7-22 (2022) [40], the generalized extreme value (GEV) distribution is employed to model G_r at each location. Apart from the GEV distribution, other distributions such as the log-normal [44,47–50] and gamma distribution [51] have been employed to fit G_l. The ultimate objective is to select a distribution that effectively describes the entire history of the annual maximum SWE, with a special focus on the upper tail of the distribution. The major issue in estimating the probability distribution associated with G_l is the fact that the observations of G_l are often estimated from observations of annual maximum snow depth. These estimates are known to contain less variability than direct measurements of the annual maximum SWE. Section 3 demonstrates our attempt to mitigate the bias in the scale parameter estimates of the site-specific probability distributions associated with G_l.

3. Methods

The appeal of the bootstrap method, originally introduced by Efron (1979) [52], is in its wide applicability to complex data structures in both parametric and non-parametric problems, as pointed out by Diciccio and Romano (1988) [53]. Both Davison and Hinkley (1997) [54] and Hall (1992) [55] demonstrate the use of the bootstrap method to calculate confidence intervals and conduct hypothesis testing when the theoretical distribution is unknown. They also discuss the utilization of the bootstrap method for bias correction in parameters, as well as for parametric and non-parametric regression when asymptotic assumptions are unfulfilled. See Hall (1992) [55] for a review of the asymptotic properties and theory underpinning the bootstrap method. There exist many examples across several domains of the use of the bootstrap method to correct for bias. For instance, the bootstrap technique has been employed in time series [56–59], dynamic panel models [60,61], and regression models [62] to rectify bias in various estimators. It has also been used in machine learning models to minimize bias in the predicted response variable [25] and in the maximum likelihood estimation (MLE) [63]. These applications, among others, illustrate the efficacy of the bootstrap method in mitigating bias. There are potential risks associated with using bias correction techniques, such as increased variability in the estimate that could outweigh the reduction in bias, leading to an increase in the prediction error [61]. Despite this, the bootstrap method can prove valuable in instances where an analytical bias correction is either unavailable or deemed inappropriate.

In Section 1, we demonstrated that the use of a regression model for predictions inherently underestimates the variance of the true response variable. This underestimation of variance subsequently influences distribution parameter estimation when the distribution is estimated using predicted values [6]. In this paper, both standard deviation and scale will be represented using the symbol σ. This section utilizes residual bootstrapping as a method to rectify the underestimated second-order (scale) parameter in the distribution of interest. The use of the residual bootstrapping is beneficial for rectifying this bias as it allows us to account for the neglected term in Equation (2).

We first make use of the regression model described in Equation (1). The function f can be a parametric or non-parametric function of the predictor variable matrix $\boldsymbol{X}$. In order to correct for the bias in the distribution parameter (σ) of $\boldsymbol{Y}$, we construct multiple

distribution parameters over B bootstrap replicates using a residual bootstrap according to the following algorithm:

1. Train/fit $f(\boldsymbol{X})$ using available data.
2. Obtain residuals $\boldsymbol{\epsilon} = \boldsymbol{Y} - f(\hat{\boldsymbol{X}})$, which are assumed to follow a normal distribution.
3. Fit the true Y_i to a specified distribution and obtain the initial scale parameter of the distribution σ^0.
4. Letting b represent one of B bootstrap samples, perform the following:
   - Obtain new responses by bootstrapping the residuals and adding to the predicted responses:
   $$\boldsymbol{Y^b} = f(\hat{\boldsymbol{X}}) + \boldsymbol{\epsilon}^b \tag{9}$$
   where $\boldsymbol{Y}^b$ represents a simulated alternative to $\boldsymbol{Y}$ based on a bootstrap sample of residual terms.
   - Estimate the scale parameter of the specified distribution (σ^b) and other related parameters (i.e., μ^b, ξ^b) for each boostrap set ($\boldsymbol{Y}^b$).
5. Find the percentile (p^*) of the estimates of $\sigma^1, ..., \sigma^B$ that is closest to $\boldsymbol{\sigma^0}$. The closest spread parameter is denoted as σ^*.
   - Sort the boostrap estimates $\sigma^1, ..., \sigma^B$ in ascending order.
   - For each estimate σ^i (where $i = 1, 2, ..., B$), calculate the absolute difference from σ^0:
   $$\Delta_i = |\sigma^i - \sigma^0| \tag{10}$$
   - Find the estimate σ^* that has the smallest absolute difference from σ^0. Mathematically, this is equivalent to finding the index j such that $j = \arg\min_i |\Delta_i|$. Then, $\sigma^* = \sigma^j$, and σ^* is accompanied by its corresponding parameters (i.e., μ^j, ξ^j).
   - Calculate the percentile position of σ^* within the sorted list of estimates. The percentile p^* is given by the following formula:
   $$p^* = \frac{\text{position of } \sigma^* \text{ in the sorted list}}{B} \times 100\% \tag{11}$$

In essence, the primary objective of the algorithm outlined above is to learn a specific percentile (p^*) from the training dataset. Subsequently, when presented with a bootstrap sampling distribution of sigma ($\sigma^1, ..., \sigma^B$) from a test dataset, we calculate the adjusted σ^* by applying the learned percentile (p^*) obtained from the training data. Obtaining the σ^* can be achieved by the following using test data when the response variable is predicted:

1. Predict $\hat{\boldsymbol{Y}}$ using the regression-based model in the above algorithm.
2. Letting b represent one of B bootstrap samples, perform the following:
   - Fit $\hat{\boldsymbol{Y}}$ to a specified distribution and obtain σ^b and other corresponding parameters (i.e., μ^b, ξ^b).
3. Calculate the percentile position of sigma ($\sigma^1, ..., \sigma^B$) within the sorted list of estimates. The percentile p^b is given by the following formula:
   $$p^b = \frac{\text{position of } \sigma^b \text{ in the sorted list}}{B} \times 100\% \tag{12}$$
4. Given a specific statistic value p^* from the above algorithm, find the σ^* in the bootstrap distribution that is associated with p^*. Conceptually, this can be expressed as $\sigma^* = \sigma^b$ for the b where $|p^b - p^*|$ is minimized. The computed σ^* is accompanied by its corresponding parameters (i.e., μ^b, ξ^b).

In the context of spatial analysis, the given algorithm primarily corrects the bias of the σ parameter for single-location data but can be extended to unobserved locations using the p^* percentile from nearby sites. This is vital where the response variable Y_i is entirely

unavailable for a particular area. Utilizing data from available locations enables bias adjustment in the σ parameters of unobserved sites. The algorithm adapts to multiple locations through a unified regression model and bootstrap correction at each site. An aggregated p^* value, either a global average or a regional one, can be derived from multiple locations for bias correction in unobserved areas.

4. Simulations

In this section, we conduct simulation experiments to assess the efficacy of our distribution-scale bias correction process. Our proposed method is evaluated on two different data-generating models, each with sample sizes of 10,000 and 1000 with the following form:

$$Y_i = \beta_0 + \sum_{j=1}^{r} \beta_j X_{ij} + \epsilon_i \tag{13}$$

Here, Y_i represents the response variable for the i-th observation, where $i = 1, ..., n$ or the sample size (n = 10,000 or 1000). The term X_{ij} represents the j-th predictor variable for the i-th observation, where $j = 1, ..., r$. β_j is the coefficients for the predictor variables, β_0 is the intercept term, and ϵ_i represents the error term for the i-th observation.

For each model, the simulation is run 200 times. In each run, the observations are partitioned into a 70/30 training/test split. The first model is a Gaussian model where the following apply:

- $X_{ij} \sim \mathcal{N}(\mu_j, 1.8)$, with $\mu_j \sim \mathcal{N}(5, 5)$ for $j = 1, \ldots, 10$;
- $0 \leq \text{Cov}(X_{ij}, X_{ik}) \leq 0.5$ for $i = 1, \ldots, n$ and $j, k = 1, \ldots, 10$ where $j \neq k$;
- $\beta_0 = 0$ and $\beta_j = 1$ for $j = 1, \ldots, 10$;
- $\epsilon_i \sim \mathcal{N}(0, 2)$ for $i = 1, \ldots, n$.

The above conditions define the structure of the simulated data for the Gaussian model where each predictor variable X_{ij} follows a normal distribution with mean μ_j and a variance of 1.8. The mean μ_j for each predictor is itself drawn from a normal distribution. This setup ensures that each predictor variable X_{ij} has some inherent variability, and this variability differs independently between predictors.

The second model is a uniform model with a linear structure similar to the form described by Equation (13) where the following apply:

- $X_{ij} \sim \text{U}(-1, 1)$ for $j = 1, \ldots, 5$;
- $\beta_0 = -5$, $\beta_1 = 4$, $\beta_2 = 2.5$, $\beta_3 = 5$, $\beta_4 = 2$, and $\beta_5 = 1$;
- $\boldsymbol{\epsilon_i} \sim \mathcal{N}(0, 2)$ for $i = 1, \ldots, n$.

The uniform model specifies that each predictor variable X_{ij} for the i-th observation and the j-th observation predictor is drawn from a uniform distribution over the interval $[-1, 1]$. The inclusion of a uniform model alongside the Gaussian model in the simulation serves the purpose of investigating the performance and scale/variance parameter behavior of different models when data follow varying distribution assumptions.

4.1. Experimental Setup

To set up our experiment, we begin by training separate linear regression models for the two simulated training datasets. We estimate the response variable from the respective regression models and fit the response variable estimates to a normal distribution. Next, we apply our proposed method to correct for the bias in σ. Note that we adjust σ here because it is the parameter that is systematically underestimated in our application of interest. This is achieved by the following:

- Selecting the best percentile (p^*) from the sampling distribution of the σ parameter that corrects for the bias in σ for the training data.
- Generating the sampling distribution of μ and σ using the test dataset with a bootstrap sample size of 200.

- Using p^*, selecting the least biased σ parameter along with its μ parameter from the sampling distribution for the test data.

The above process describes the bias correction for a set of data observations. This process is applied to the 200 data simulations for the Gaussian and uniform models.

4.2. Results

The results of applying our bias correction method for estimating distribution parameters are presented in Figure 2. The degree of bias is assessed through the ratio of estimated to true parameters. A ratio of one implies no bias, whereas a ratio below one signifies an underestimation, and a ratio above one signifies an overestimation. The second quadrant (set of boxplots) of the figure reveals that "unadjusted" parameters consistently underestimate the value of σ when the data generation sample size is 10,000. Conversely, the first quadrant shows that our bias correction method effectively mitigates this bias in σ (at n = 10,000), albeit at the expense of a minor increase in the variance of the ratios for μ. This pattern—bias elimination in σ and a slight increase in variance for μ—is also evident in the third and fourth quadrants, which are associated with a data generation sample size of 1000. These results demonstrate that our technique can achieve some form of bias correction, regardless of the sample size.

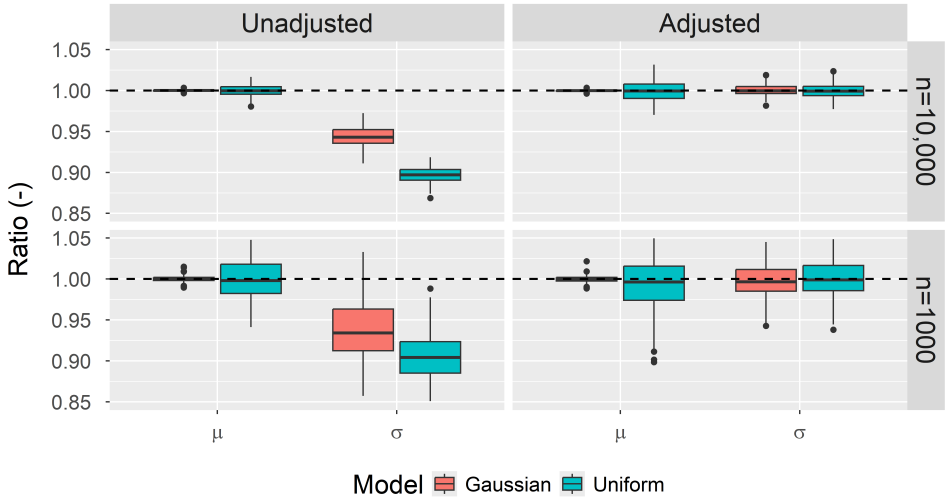

Figure 2. Boxplots comparing the ratio of the normal distribution parameters (mean and scale) of the actual response variable versus the predicted response variable. The data generation models, i.e., the uniform and Gaussian models, are generated at different sample sizes. Ratios are shown both before (unadjusted) and after (adjusted) the bias correction technique is employed. A ratio closer to one means that the parameter is less biased, while ratios below one indicate an underestimate and vice versa.

Table 2 quantifies the bias and coefficient of variation (COV) for the unadjusted and adjusted parameters at different sample sizes. The quantities' bias and COV are computed by comparing the estimated parameter with the true parameter, utilizing a relative ratio approach for each parameter as shown in Figure 2. On the other hand, the COV quantifies the relative variability or dispersion of the ratio of the parameter compared to its mean. A higher COV value indicates a larger amount of variability relative to the mean and vice versa. Table 2 demonstrates results that align with the visual findings presented in Figure 2. Specifically for bias across different sample sizes, both the uniform and Gaussian models show that μ is estimated without bias—evidenced by a ratio value of 1—regardless of whether bias correction is applied. Conversely, σ is systematically underestimated in the absence of bias correction, registering ratio values of 0.9 and 0.94 for the uniform and Gaussian models, respectively. Importantly, the application of our bias correction technique rectifies this underestimation, bringing the average bias value for σ up to 1 across

different sample sizes. Regarding the COV, the table reveals that a reduction in sample size is associated with increased variability. In both the uniform and Gaussian models, decreasing the sample size leads to an increase in the COV for both μ and σ, regardless of whether the parameters are adjusted. In summary, the results in Table 2 demonstrate that the bias correction method effectively resolves the underestimation issue observed in the unadjusted parameters.

Table 2. Comparison of the bias and coefficient of variation (COV) of unadjusted and adjusted parameters for the uniform and Gaussian models generated at different sample sizes. The quantities are computed based on the relative ratio with respect to the true parameter values.

		Uniform Model				**Gaussian Model**			
		COV		**Bias**		**COV**		**Bias**	
		μ	σ	μ	σ	μ	σ	μ	σ
n = 10,000	Unadjusted	0.01	0.01	1.00	0.90	0.00	0.01	1.00	0.94
	Adjusted	0.01	0.01	1.00	1.00	0.00	0.01	1.00	1.00
n = 1000	Unadjusted	0.04	0.04	1.00	0.90	0.00	0.04	1.00	0.94
	Adjusted	0.03	0.03	1.00	1.00	0.00	0.02	1.00	1.00

5. Applications

Section 1 describes the inevitable underestimation of true SWE variance that comes when using estimates of the SWE derived from snow depth measurements. In this section, we demonstrate how our bootstrap bias correction approach can be used to mitigate this issue. The data used in this demonstration come from Wheeler et al. (2022) [6], who proposed an RF model to estimate the annual maximum snow load (or, equivalently, the SWE) from the annual maximum snow depth. The dataset includes annual maximum pairs of SWE/snow depth from the Global Historical Climatology Network—Daily (GHCND) [64]. Note that the maximum snow depth and maximum SWE need not occur on the same day of the snow season. The sources of these measurements include both National Weather Service First-Order Stations (FOS) and Natural Resource Conservation Service Snowpack Telemetry (SNOTEL) stations [64]. FOS are typically situated at airports and are one of the few sources of direct measurements of the SWE in the GHCND outside of SNOTEL stations. Additionally, SNOTEL stations are automated sites located in remote, high-elevation mountain watersheds in the western U.S. The consolidated dataset also includes Snow Course (SC) observations from the Natural Resources Conservation Service [65]. SC stations typically consist of manual measurements of the SWE taken on a monthly basis. Additionally, the dataset also includes region-specific snow measurements from Maine (ME) [66] and New York (NY) [67]. Figure 3 shows a map of these data sources, which are available for download from Wheeler (2021) [68].

Table 3 describes all the variables used in the model development phase, including the ancillary variables (such as temperature and distance to the coast) used in the RF model building phase. The data span the period 1915–2020, with 45,879 observations from 2473 stations over a period of 105 years, though not all stations were active for all 105 seasons. Of the 45,879 observations, a total of 25,523 are used to train the model, with 22,356 observations used for validation.

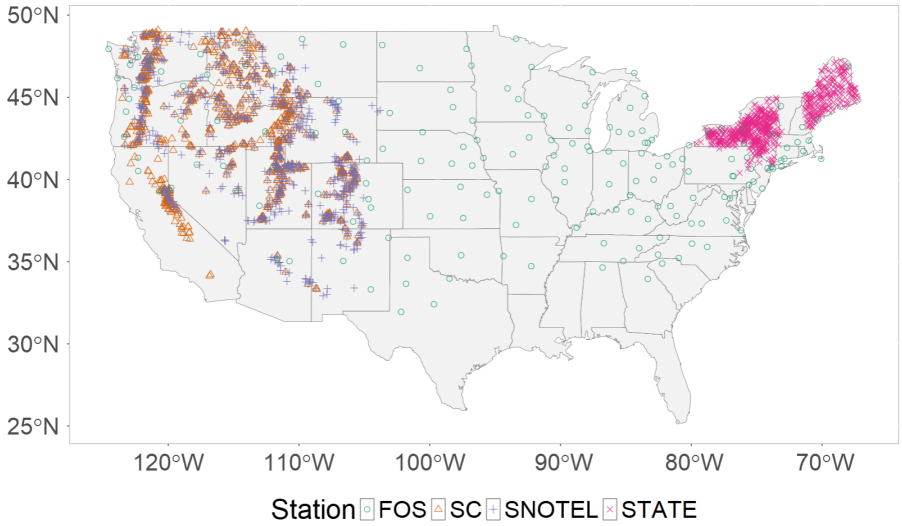

Figure 3. Geographic locations of the snow measurement stations described in Wheeler et al. (2022) [6]. Station types include First-Order Stations (FOS), Snow Course (SC), and Snowpack Telemetry (SNOTEL) stations, as well as measurements from two state-specific data sources (STATE) from New York and Maine.

Table 3. Description of variables used in the model development phase as originally presented in Wheeler (2021) [6]. The data range is from 1915 to 2020.

Climate Variables			
Name	**Description**	**Units**	**Variable**
SD	Snow Depth	mm	h_t
SWE	Snow Water Equivalent	mm	s
MCMT	Mean Coldest Month Temperature	°C	T_c
MWMT	Mean Warmest Month Temperature	°C	T_w
TD	MWMT–MCMT	°C	T_d
PPTWT	Sum of Winter Precipitation (December–February)		P_t
Non-Climate Variables			
D2C	Distance to Coast	m	D_c
ELEV	Elevation	m	E
SMONTH	Month of Max Depth (1 October, 9 June)		M_s

5.1. Depth-to-Load Conversion Model

Using these data, we build an RF depth-to-load conversion model similar to the model built by Wheeler et al. (2022) [6]. In addition, we build gradient boosting machine (GBM) and support vector regression (SVR) depth-to-load conversion models. The depth-to-load models are represented by Equation (14), where a prediction of snow water content ratio (η_{model}) is multiplied by snow depth (h_t) to estimate the SWE.

$$SWE_t = h_t * \eta_{model}(h_t, T_c, T_w, T_d, P_t, D_c, E, M_s) + \epsilon_t \quad (14)$$

The response variable for the snow density model is a ratio of the SWE to snow depth ($\frac{s}{h_t}$), which is referred to hereafter as the specific gravity of snow. In the context of SWE modeling, estimating the specific gravity of snow is favored over direct estimates of the SWE, as described by Sturm et al. (2010) [4]. To evaluate the impact of the conversion models on the design snow load, we assume that all measurement locations have annual maximum SWE values that follow a log-normal distribution, though other extreme value distributions could replace the log-normal distribution without loss of generality.

We estimate the parameters of the log-normal distribution using both direct measurements of the SWE and the estimated measurements of the SWE from the three previously

described machine learning models. For each location, we estimate the distribution parameters using maximum likelihood estimation. Figure 4 displays a boxplot of the estimated parameter ratios, specifically, $\frac{\mu^*}{\mu}$ and $\frac{\sigma^*}{\sigma}$. Table 4 shows the count of weather stations per weather station network. Note that we take the exponent of the log-normal parameter in order to make comparisons on the original scale of the data. This makes the final results easier to interpret and emphasizes the practical differences between approaches. For computing the parameter ratios, μ^* and σ^* correspond to the distribution parameters for the observed loads, while μ and σ denote those for the estimated loads according to the depth-to-load conversion method utilized in the RF model. The spread parameter, our parameter of interest, tends to be consistently undervalued for the FOS, ME, NY, and SNOTEL weather station networks, as illustrated in Figure 4. The pattern is replicated in the distribution of estimated loads when we employ the SVR and GBM algorithms, as shown in Figure 5. Considering that σ plays a significant role in estimating design ground snow loads, an underestimation of the log-normal distribution variance will ultimately result in underestimations of the true design loads.

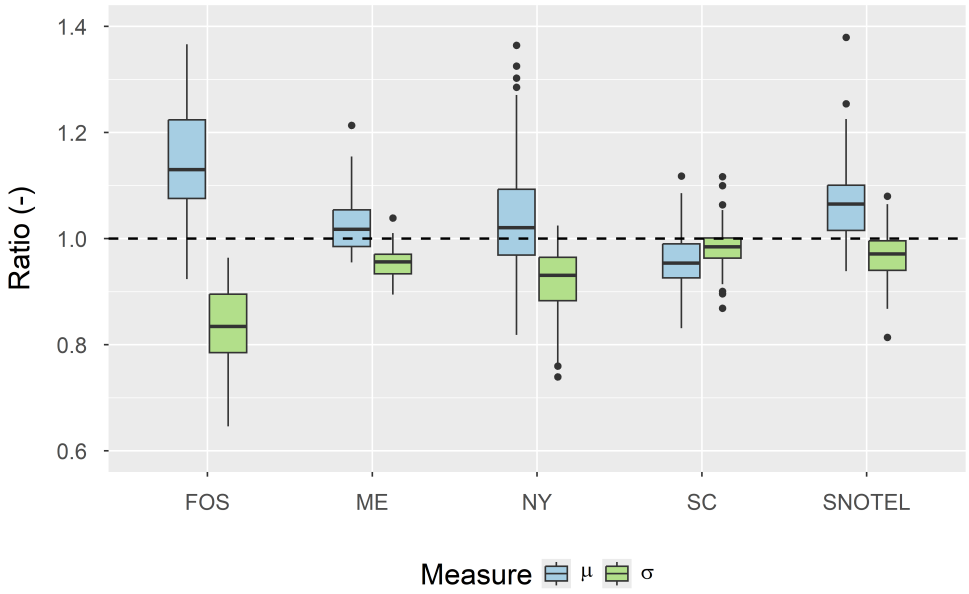

Figure 4. Boxplots comparing the ratio of the log-normal distribution parameters (mean and scale) of the actual snow load to the predicted snow load from the RF model before any bias correction. A ratio closer to one means that the parameter is less biased, while ratios below one indicate an underestimate and vice versa. Distributions are fitted to each weather station with the sample size of each network provided in Table 4.

Table 4. Number of qualified ($n \geq 20$) weather stations within each weather station network.

Network	Name	Number of Weather Stations
FOS	First-Order Stations	46
ME	Maine State Network	23
NY	New York State Network	122
SC	Snow Course	144
SNOTEL	Snowpack Telemetry	74

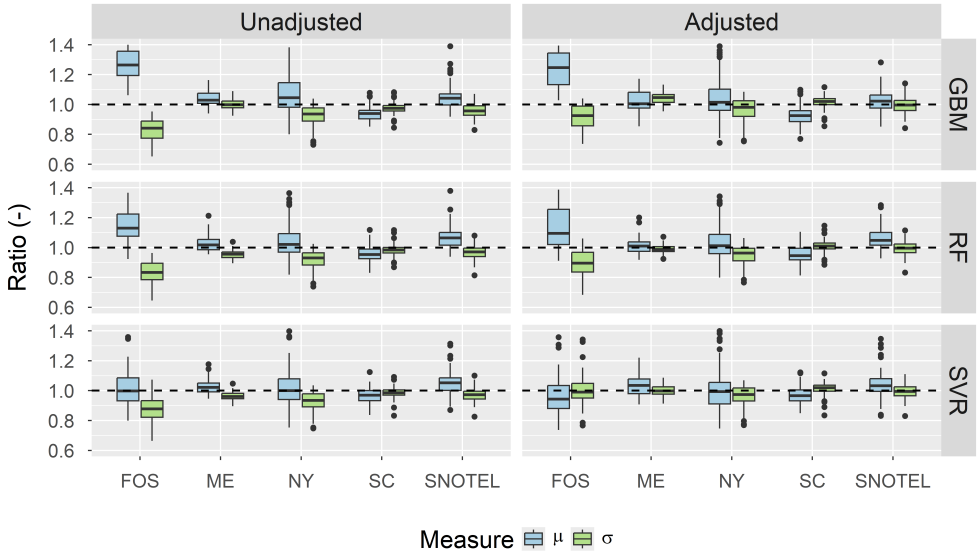

Figure 5. Boxplots comparing the ratio of the log-normal distribution parameters (mean and scale) of the actual snow load versus the predicted snow load, before and after bias correction. Ratios are shown both before (unadjusted) and after (adjusted) the bias correction technique is employed. A ratio closer to one means that the parameter is less biased, while ratios below one indicate an underestimate and vice versa. Distributions are fitted to each weather station with the sample size of each station cohort provided in Table 4. Predicted snow loads are estimated using the gradient boosting machine (GBM), random forest (RF), and support vector regression (SVR).

5.2. Bias Correction in Snow Load Distribution

Figure 5 shows the results of our proposed bias correction method, described in Section 3, when applied to the different machine learning models. To achieve this, we first train conversion models to predict the SWE and obtain the error distribution of the respective conversion models. Subsequently, we apply the bootstrap process to each weather station, defining p^* at the station level as the optimal percentile for a sampling distribution of parameters that minimizes bias in the variance parameter. Only weather stations with 20 or more joint annual observations of snow depth and the SWE are used to obtain estimates of p^*. We recognize that, even with 20 observations, the parameter estimates may be highly uncertain. Despite this limitation, there is still value in conducting comparisons across different station cohorts (FOS, ME, NY, SC, SNOTEL) to identify potential systematic biases, a task that becomes unfeasible if the minimum observation threshold increases. Our training data comprise 211 distinct weather stations that are separate from the 211 stations used in the test data. For each weather station in the training data, we use 500 bootstrap samples to estimate several sets of distribution parameter estimates, from which we select the p^* for our estimates σ that most closely correspond with the σ estimates obtained from direct measurements of the SWE.

Next, SWE distribution parameter predictions are made for the test data. Along with the error distribution from the conversion model, Step 3 of the algorithm described in Section 3 is applied using 500 bootstraps to obtain a sampling distribution of the log-normal distribution parameters. Given that p^* is a station-specific calculation, we compute the global selection parameter (p^*_α), which is 0.67, an average of the p^* values. The parameters of the log-normal distribution displayed in Figure 5 are the parameters before and after the bias correction of the parameters across various machine learning models. Bias is assessed by comparing the ratio of the estimated parameter (converted load) to the true parameter (direct load). Prior to the correction of bias, a majority of the σ values across the different station cohorts for each machine learning model are found to be underestimated. In contrast, the mean parameter values (μ) typically exhibit less bias. However, there is a notable exception for the mean values (μ) in the case of the GBM and RF models in the

First-Order Station cohort. The "adjusted" portion of Figure 5 depicts the parameters' post-bias correction. The plot reveals that the scale parameters across the various station networks align more closely with the reference line, indicative of a successful bias correction. In particular, the SVR model exhibits significant improvement, effectively mitigating the bias in σ for various station cohorts after adjustment. Conversely, while the GBM and RF models demonstrate some degree of bias correction following the adjustment, they fall short of fully correcting the bias either for σ or μ at some station networks. This shortfall is especially noticeable in the First-Order Station (FOS) cohort. In conclusion, the efficacy of our bias correction strategy is contingent upon the type of model employed. As observed, a better model like the SVR facilitates full bias correction, in contrast to relatively less effective models such as the GBM and RF models (tree-based models), which do not fully achieve the desired bias correction. This discrepancy can be attributed to the inability of tree-based models to extrapolate predictions beyond the range of observed data. This leads to a tendency for the predictions to overestimate lower extremes and underestimate higher extremes, and our results show that bias in predictions for the extremes cannot be overcome with our bias correction method alone. This highlights the important role of model choice in the successful implementation of our bias correction approach.

We also evaluate our bootstrap bias correction methodology using the GEV distribution across diverse station cohorts for each machine learning model. The process for bias correction remains identical, the only difference being the utilization of the GEV distribution for distribution estimation instead of the log-normal distribution. In the context of the GEV distribution, $\tilde{\mu}$ and $\tilde{\sigma}$ are used to denote the location and scale parameters, respectively. Figure 6 shows the "unadjusted" and "adjusted" parameters when the different machine learning models are used. Just like in Figure 5, bias is measured as a ratio of the estimated parameter (converted load) to the true parameter (direct load). Before bias correction, the unadjusted scale parameter for the GEV is less biased compared to its counterpart in the log-normal distribution in Figure 5. Nevertheless, following the bias correction, we see that the bias in the scale parameter ($\tilde{\sigma}$) aligns more closely with the reference line, demonstrating a successful adjustment for all station types except SNOTEL stations. Fortunately, SNOTEL stations are the lone station group with complete SWE measurements, which means our adjustment approach would not be required for that station type.

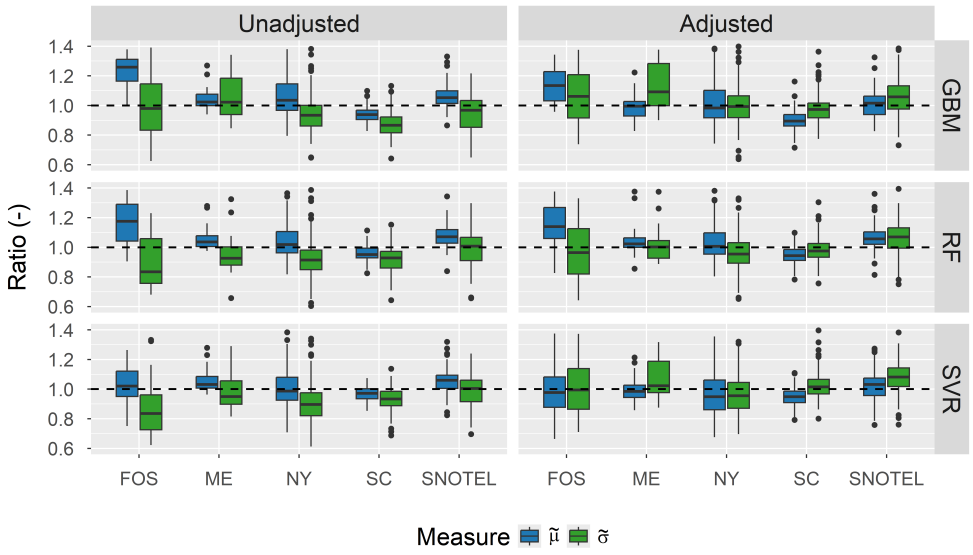

Figure 6. Boxplots comparing the ratio of the GEV distribution parameters (location and scale) of the actual snow load versus the predicted snow load, before and after bias correction. Ratios are shown both before (unadjusted) and after (adjusted) the bias correction technique is employed. A ratio closer to one means that the parameter is less biased, while ratios below one indicate an underestimate and vice versa. The distribution is fitted on the test dataset per weather station. Distributions are fit to each

weather station with the sample size of each station cohort provided in Table 4. Predicted snow loads are estimated using the gradient boosting machine (GBM), random forest (RF), and support vector regression (SVR).

On the other hand, Figure 7 shows the change in the shape parameter across the different station type groups. Comparing the "adjusted" to the "unadjusted", we observe slight increases in the shape parameter when the bootstrap method is applied. The slight increase suggests that the proposed bootstrap approach, which primarily focuses on the scale parameter, makes no appreciable difference to the ξ parameter estimates. This outcome aligns with our expectations since the approach specifically targets the scale parameter rather than the ξ parameter.

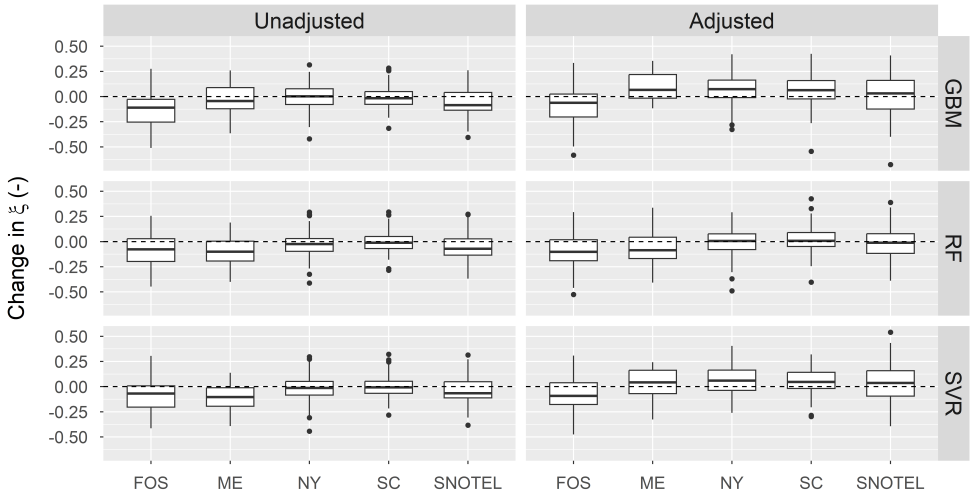

Figure 7. Boxplots comparing the raw change in the shape parameter of the GEV distribution for the actual snow load versus the predicted snow load, before and after bias correction. Ratios are shown both before (unadjusted) and after (adjusted) the bias correction technique is employed. A change closer to zero means that the parameter is less biased, while a change below zero indicates an underestimate and vice versa. Distributions are fit to each weather station with the sample size of each station cohort provided in Table 4. Predicted snow loads are estimated using the gradient boosting machine (GBM), random forest (RF), and support vector regression (SVR).

5.3. Reliability-Targeted Load (RTL) Results after Bias Correction

In this section, we assess the impact of our bias correction on the reliability-targeted loads (RTLs) calculated in [46]. RTLs, as explained in Section 2.2, account for both the variability in environmental hazard (e.g., snow) and the variability in structural resistance to the hazard via Monte Carlo simulation. The computation of RTLs is based on the same assumptions presented by Bean et al. (2021) [46], with the only difference being in the manner in which the ground snow load probability distribution parameters are calculated. We employ our distribution-scale bias correction method to calculate the probability distribution parameters with the objective of mitigating any bias in the scale parameter when estimating missing snow load values. In this process, we depend on the global estimate of $p^* = 0.67$, whose value was derived in the previous subsection, for bias correction of the scale parameter. Bootstrap samples of $B = 500$ are repeated for the 6727 Tier 1 and 510 Tier 2 stations described by Bean et al. (2021) [46]. Tier 3 stations from that report are excluded from this analysis due to their insufficient number of non-zero observations, which renders them unsuitable for our bias correction approach. Note that some stations had years with direct (observed SWE) and years with indirect (i.e., estimated SWE) measurements of snow load. In such cases, bootstrap sampling is only applied

to indirect measurements of snow load, and direct measurements of snow load remain identical in each iteration of the bootstrap sampling.

Figure 8 visually represents the percentage change in RTLs that results when using the bias-corrected probability distribution parameters. When examining the number of stations with a positive increase in RTL out of 7237 stations, we observe 81%, 82%, 82%, and 82% positive changes in RTL for RT_I, RT_II, RT_III, and RT_IV, respectively. Focusing on the interquartile, we observe an approximate 10% to 13% median percent change in RTL values across the risk categories. It is important to highlight that, while we typically observe a 10% to 13% median increase in RTLs, there is a very large range of RTL relative changes evident within the boxplots. Among these outliers, some stations experience substantial reductions in RTLs. These stations seem to be primarily located in the southern parts of the CONUS, and the precise cause of these anomalous observations will require explorations beyond the scope of this paper. Conversely, there are also substantial increases in RTLs observed at some locations. These large differences highlight the sensitivity of the distribution fitting process for RTLs at measurement locations with small sample sizes. These findings also shed light on one of the limitations inherent in utilizing a global percentile estimate for the scale parameter adjustment. One alternative approach would be to use a local or regional adjustment percentile, though this would require a more spatially comprehensive network of co-located measurements of the snow depth and SWE in order to estimate the bias correction percentile precisely. This would be challenging as most weather stations with co-located depth and SWE measurements are predominantly situated in mountainous regions.

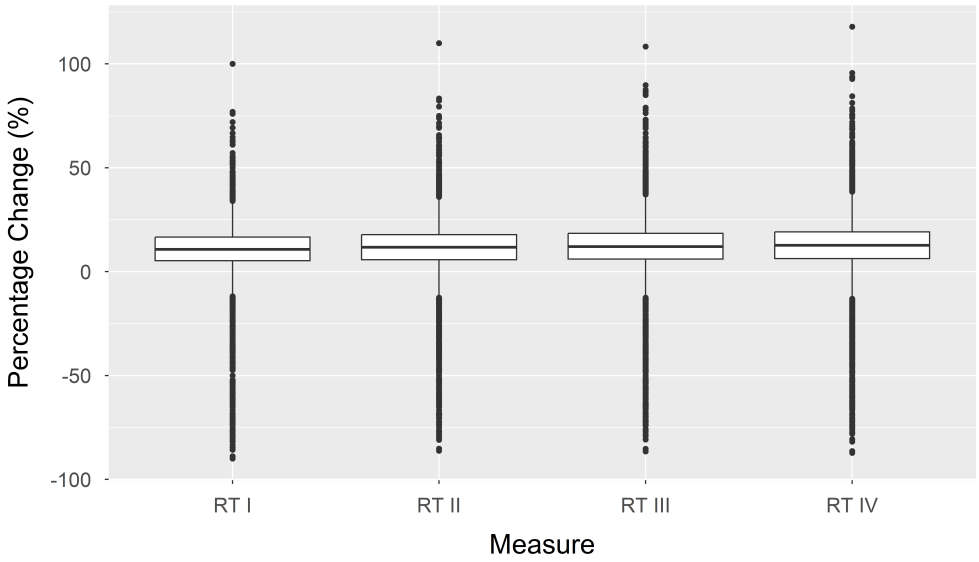

Figure 8. Boxplot comparing changes in reliability-targeted loads (RTL) after bias correction for each risk category. RT I, RT II, RT III, and RT IV are associated with a risk index of 2.5, 3, 3.25, and 3.5, respectively. The RTLs are computed using the same data and assumptions as used by Bean et al. [46].

5.4. Limitations

There are limitations to our bias correction method worth considering before applying it to other problems. Firstly, the success of our method in reducing bias hinges on the quality of the regression model used to generate the indirect observations. In this context, a good regression model is one that produces unbiased predictions of the mean response. If, however, a model has biased predictions of the mean response, the bootstrap approach to correcting the variance in the distribution of estimates will fail. Thus, it is crucial to opt for a model that, at a minimum, provides unbiased predictions of the mean response, even if the variance of collections of those predictions is biased, to ensure the effectiveness of our bias correction method.

To demonstrate the effect of an inadequate model on bias correction, we apply an RF model to the Gaussian simulation problem in Section 4. Zhang et al. (2019) [69] highlights the fact that RF models may struggle to represent highly linear relationships. This can result in challenges in representing the true linear relationship with the RF model. Figure 9 illustrates a single fit of the RF model applied to the linear problem formulated in the Gaussian version of Equation (13). The model struggles to fit the observations on the edges of the explanatory variable space, resulting in inconsistent errors across different levels of the response variable. Figure 10 shows that, despite the attempt to correct the bias in the scale parameter (σ), the RF model's inefficiency limits our proposed approach's effectiveness. In other words, we are unable to achieve a complete correction because our bootstrap technique cannot overcome the bias introduced by the structural deficiencies in the RF prediction on the edges of the predictor variable space.

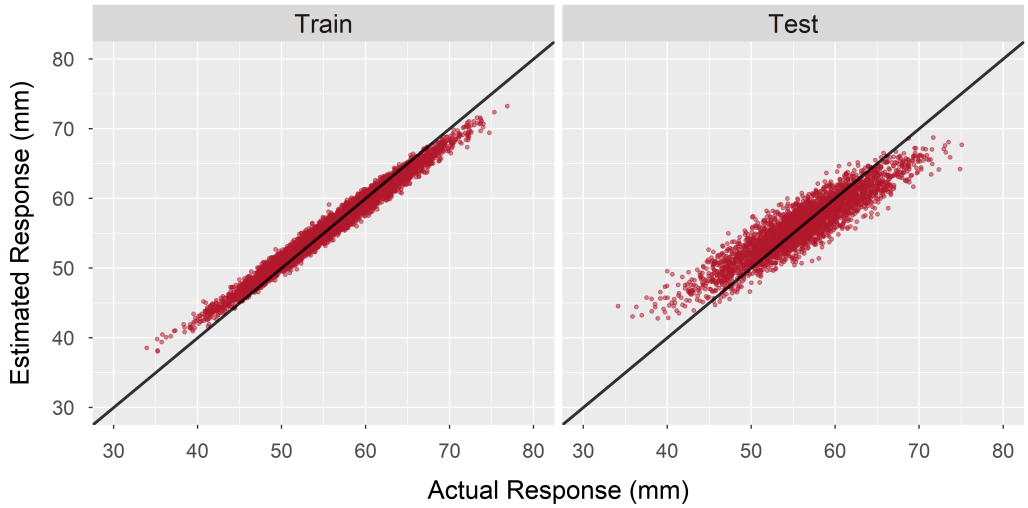

Figure 9. Scatter plot comparing the actual versus estimated response variable for a linear problem using a random forest model.

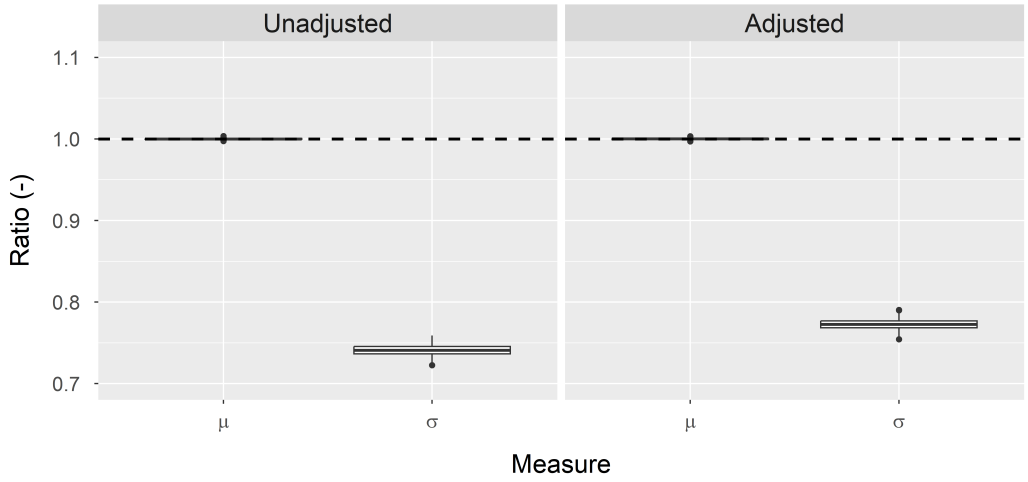

Figure 10. Boxplot comparing the ratio of the normal distribution parameters of the actual vs. estimated response variable when bias is corrected or not for a linear problem. The response variable is estimated using a random forest model. Ratios are shown both before (unadjusted) and after (adjusted) the bias correction technique is employed. A ratio closer to one means that the parameter is less biased, while ratios below one indicate an underestimate, and ratios above one indicate an overestimate.

Additionally, Figure 5 highlights another limitation of the bootstrap method. When the variance/scale parameter (σ) is already unbiased, such as is the case with the Snow Course (SC) weather station network, the application of our approach tends to introduce an overestimation bias in parameter estimation. While overestimated parameters are generally preferable to underestimated parameters for the purposes of engineering design, caution must be exercised when interpreting the results and applying them in practical design scenarios.

Note that the alternative CDF matching described by Cho and Jacobs (2020) [8] has its own limitations. As detailed in Section 2.1, this alternative bias corrects UA SWE using SNODAS SWE through CDF matching. The impact of this alternative correction is depicted in Figure 11. For this result, we obtain the estimated parameter from the log-normal distribution of the corrected UA SWE, while the true parameter is derived from the log-normal distribution of the observed SWE. The bias is quantified by calculating the ratio between the estimated parameter and the true parameter. If we do not correct for bias, the first set of boxplots in Figure 11 illustrates that the unadjusted parameters result in some systematic underestimation of σ for the FOS, ME, and NY weather station types. However, when applying the proposed bias correction method, the bias in σ for the adjusted parameters becomes worse for the FOS, ME, and NY weather station types. An important future task is to explore to what degree the limitations in both approaches are being exaggerated by misreported observations of snow depth/SWE within each station network.

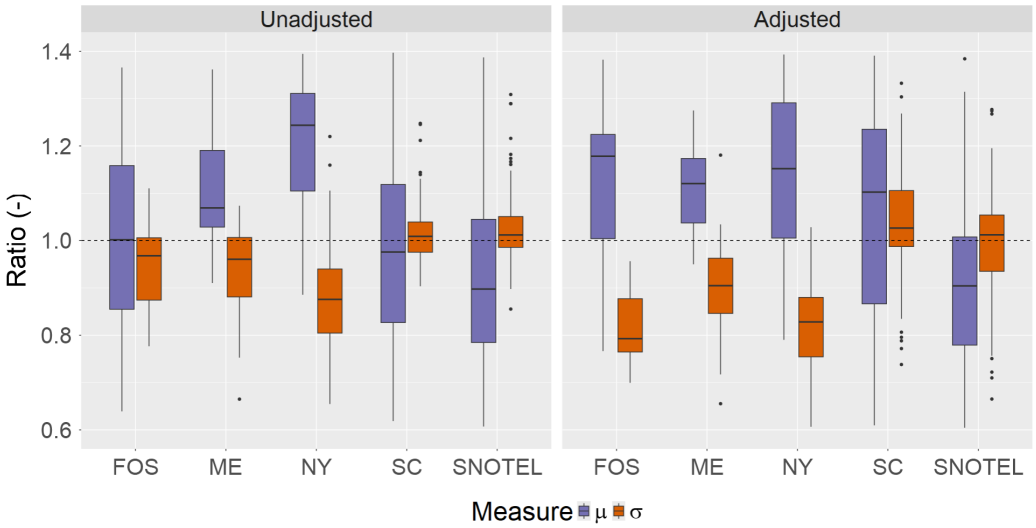

Figure 11. Boxplot comparing the estimated log-normal distribution parameters (μ and σ) using the UA SWE versus ground SWE before and after bias correction. A ratio closer to one means that the parameter is less biased, while ratios below one indicate an underestimate and vice versa.

6. Conclusions

Although regression-based SWE estimation models significantly expand the number of weather stations available for future mapping efforts, they come with the drawback of consistently underestimating the variances in the resulting distributions. This underestimation has significant implications for the design and safety of structures, particularly in the Conterminous United States (CONUS), where extreme snow events can pose a risk. To address this, our study introduced a distribution-scale bias correction technique to rectify the variance and/or scale parameter biases in SWE distribution. This approach is based on residual bootstrapping and effectively reincorporates the scale/variance bias back into the distribution fitting process. We evaluated the performance of our method using both synthetic and actual datasets. In synthetic data, the method effectively corrected the underestimated scale parameter arising from regression model estimates. In the case of snow data, the technique rectified the scale parameter bias without affecting other

non-scale distribution parameters across multiple types of network stations (FOS, ME, NY, SC, and SNOTEL). This suggests that stations not directly measuring the SWE can still yield unbiased scale parameters when the SWE is estimated through a regression-based approach.

Although our proposed technique successfully addresses the problem of underestimating variance, it comes with its own set of challenges. One issue is that the efficacy of our method is contingent upon choosing the correct type of regression model. Moreover, we observed that applying our bias correction method to situations where the scale parameter is already unbiased can lead to its overestimation. This implies that our technique should only be used when there is clear evidence of bias in the scale parameter. Presently, our approach employs a global selection parameter for identifying unbiased scale parameters, but future work could explore the use of local-level selection criteria. We expect that adopting a local-level parameter could refine the identification of unbiased scale parameters. Despite these limitations, our bias correction approach offers a valuable tool for rectifying scale parameter bias in weather station data, thus enhancing the accuracy of reliability analyses that depend on SWE distribution.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/glacies1010004/s1.

Author Contributions: Conceptualization, B.B; Methodology, K.P. and B.B.; Formal analysis, K.P.; Data curation, K.P and B.B.; Writing—original draft preparation, K.P.; Writing—review and editing, B.B.; Visualization, K.P.; Supervision, B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data from the application example described in this study are available for download from Wheeler (2021) [68] and Supplementary Material.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASCE	American Society of Civil Engineers
BM	Block Maxima
COOP	Cooperative Observer Network
CONUS	Conterminous United States
CDF	Cumulative Distribution Function
ECDF	Empirical Cumulative Distribution Function
FOS	First-Order Stations
GEV	Generalized extreme value
GP	Generalized Pareto
GHCND	Global Historical Climatology Network—Daily
GBM	Gradient boosting machine
LRFD	Load And Resistance Factor Design
MAE	Mean Absolute Error
ME	Maine
MLE	Maximum likelihood estimation
MAD	Mean Absolute Deviation
MRI	Mean recurrence interval
ML	Maximum likelihood
NY	New York
NWS	National Weather Service
OSBF	One-step boosted forests

PDF	Probability Density Function
RF	Random forests
RMSE	Root Mean Square Error
RTL	Reliability-targeted load
SWE	Snow water equivalent
SNODAS	Snow Data Assimilation System
SNOTEL	Snowpack Telemetry
SVR	Support vector regression
SC	Snow Course
SNWD	Snow depth
UA	University Of Arizona
WESD	Water Equivalent of Snow on the Ground
U.S.	United States

References

1. Al-Rubaye, S.; Maguire, M.; Bean, B. Design ground snow loads: Historical perspective and state of the art. *J. Struct. Eng.* **2022**, *148*, 03122001. [CrossRef]
2. National Center for Atmospheric Research. The Climate Data Guide: ERA-Interim. 2022. Available online: https://climatedataguide.ucar.edu/climate-data/era-interim (accessed on 7 November 2022).
3. Jonas, T.; Marty, C.; Magnusson, J. Estimating the snow water equivalent from snow depth measurements in the Swiss Alps. *J. Hydrol.* **2009**, *378*, 161–167. [CrossRef]
4. Sturm, M.; Taras, B.; Liston, G.E.; Derksen, C.; Jonas, T.; Lea, J. Estimating snow water equivalent using snow depth data and climate classes. *J. Hydrometeorol.* **2010**, *11*, 1380–1394. [CrossRef]
5. McCreight, J.L.; Small, E.E. Modeling bulk density and snow water equivalent using daily snow depth observations. *Cryosphere* **2014**, *8*, 521–536. [CrossRef]
6. Wheeler, J.; Bean, B.; Maguire, M. Creating a universal depth-to-load conversion technique for the conterminous United States using random forests. *J. Cold Reg. Eng.* **2022**, *36*, 04021019. [CrossRef]
7. Ross, S. *A First Course in Probability*; Pearson: Upper Saddle River, NJ, USA, 2010; pp. 333–348.
8. Cho, E.; Jacobs, J.M. Extreme value snow water equivalent and snowmelt for infrastructure design over the contiguous United States. *Water Resour. Res.* **2020**, *56*, e2020WR028126. [CrossRef]
9. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2024.
10. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [CrossRef]
11. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
12. Kenneth, K. Distfixer: Distribution Bias Correction Using Residual Bootstrap. GitHub Repository. 2023. Available online: https://github.com/Kinekenneth48/distfixer (accessed on 7 November 2022).
13. Gilleland, E.; Katz, R.W. extRemes 2.0: An Extreme Value Analysis Package in R. *J. Stat. Softw.* **2016**, *72*, 1–39. [CrossRef]
14. Delignette-Muller, M.L.; Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *J. Stat. Softw.* **2015**, *64*, 1–34. [CrossRef]
15. Pavia, J.M. Testing Goodness-of-Fit with the Kernel Density Estimator: GoFKernel. *J. Stat. Softw. Code Snippets* **2015**, *66*, 1–27. [CrossRef]
16. R package version 3.4.2. code by Richard, A.; Becker, O.S.; version by Brownrigg, R. Enhancements by Minka, T.P.; Deckmyn, A. maps: Draw Geographical Maps. 2023. Available online: https://CRAN.R-project.org/package=maps (accessed on 27 June 2024).
17. Rinker, T.W.; Kurkiewicz, D. *Pacman: Package Management for R*; version 0.5.0.; Buffalo: New York, NY, USA, 2018. Available online: http://github.com/trinker/pacman (accessed on 7 November 2022).
18. Wright, M.N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
19. Bengtsson, H. A Unifying Framework for Parallel and Distributed Processing in R using Futures. *R J.* **2021**, *13*, 208–227. [CrossRef]
20. Ridgeway, G.; Developers, G. gbm: Generalized Boosted Regression Models; R Package Version 2.2.2. Available online: https://CRAN.R-project.org/package=gbm (accessed on 7 November 2022).
21. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab—An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]
22. Hijmans, R.J. raster: Geographic Data Analysis and Modeling; R Package Version 3.6-26. Available online: https://CRAN.R-project.org/package=raster (accessed on 7 November 2022).
23. Pebesma, E.; Bivand, R. *Spatial Data Science: With Applications in R*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2023; p. 352. [CrossRef]
24. Zhang, G.; Lu, Y. Bias-corrected random forests in regression. *J. Appl. Stat.* **2012**, *39*, 151–160. [CrossRef]
25. Hooker, G.; Mentch, L. Bootstrap bias corrections for ensemble methods. *Stat. Comput.* **2016**, *28*, 77–86. [CrossRef]

26. Breiman, L. *Using Adaptive Bagging to Debias Regressions*; Technical Report 547; Statistics Dept. UCB: Berkeley, CA, USA, 1999. Available online: https://www.stat.berkeley.edu/users/breiman/adaptbag99.pdf (accessed on 7 November 2022).
27. Ghosal, I.; Hooker, G. Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. *J. Comput. Graph. Stat.* **2020**, *30*, 493–502. [CrossRef]
28. Song, J. Bias corrections for random forest in regression using residual rotation. *J. Korean Stat. Soc.* **2015**, *44*, 321–326. [CrossRef]
29. Belitz, K.; Stackelberg, P. Evaluation of six methods for correcting bias in estimates from ensemble tree machine learning regression models. *Environ. Model. Softw.* **2021**, *139*, 105006. [CrossRef]
30. Broxton, P.D.; Dawson, N.; Zeng, X. Linking snowfall and snow accumulation to generate spatial maps of SWE and snow depth. *Earth Space Sci.* **2016**, *3*, 246–256. [CrossRef]
31. Barrett, A.P. *National Operational Hydrologic Remote Sensing Center Snow Data Assimilation System (SNODAS) Products at NSIDC*; Technical Report 11; National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences: Boulder, CO, USA, 2003. Available online: https://nsidc.org/sites/default/files/nsidc_special_report_11.pdf (accessed on 7 November 2022).
32. Carroll, T.; Cline, D.; Fall, G.; Nilsson, A.; Li, L.; Rost, A. NOHRSC operations and the simulation of snow cover properties for the coterminous US. In Proceedings of the Annual Meeting of the Western Snow Conference, Sun Valley, ID, USA, 17–19 April 2001; pp. 1–14. Available online: https://www.nohrsc.noaa.gov/technology/papers/wsc2001/wsc2001.pdf (accessed on 7 November 2022).
33. Wood, A.W.; Leung, L.R.; Sridhar, V.; Lettenmaier, D.P. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Clim. Chang.* **2004**, *62*, 189–216. [CrossRef]
34. Boé, J.; Terray, L.; Habets, F.; Martin, E. Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *Int. J. Climatol.* **2007**, *27*, 1643–1655. [CrossRef]
35. Ashfaq, M.; Bowling, L.C.; Cherkauer, K.; Pal, J.S.; Diffenbaugh, N.S. Influence of climate model biases and daily-scale temperature and precipitation events on hydrological impacts assessment: A case study of the United States. *J. Geophys. Res.* **2010**, *115*, *D14*. [CrossRef]
36. Haerter, J.O.; Hagemann, S.; Moseley, C.; Piani, C. Climate model bias correction and the role of timescales. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1065–1079. [CrossRef]
37. Piani, C.; Haerter, J.O.; Coppola, E. Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.* **2009**, *99*, 187–192. [CrossRef]
38. Ines, A.V.; Hansen, J.W. Bias correction of daily GCM rainfall for crop simulation studies. *Agric. For. Meteorol.* **2006**, *138*, 44–53. [CrossRef]
39. Lakshmanan, V.; Gilleland, E.; McGovern, A.; Tingley, M. *Machine Learning and Data Mining Approaches to Climate Science*; Springer: Berlin/Heidelberg, Germany, 2015. Available online: https://link.springer.com/book/10.1007/978-3-319-17220-0 (accessed on 7 November 2022).
40. ASCE 7-22. *Minimum Design Loads and Associated Criteria for Buildings and Other Structures*, ASCE/SEI 7-22 ed.; American Society of Civil Engineers: Reston, VA, USA, 2022. Available online: https://ascelibrary.org/doi/abs/10.1061/9780784415788 (accessed on 7 November 2022).
41. Ellingwood, B.; Galambos, T.V.; MacGregor, J.G.; Cornell, C.A. *Development of a Probability Based Load Criterion for American National Standard A58: Building Code Requirements for Minimum Design Loads in Buildings and Other Structures*; US Department of Commerce, National Bureau of Standards: Washington, DC, USA, 1980; Volume 13. Available online: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nbsspecialpublication577.pdf (accessed on 7 November 2022).
42. Ellingwood, B.; MacGregor, J.G.; Galambos, T.V.; Cornell, C.A. Probability based load criteria: Load factors and load combinations. *J. Struct. Div.* **1982**, *108*, 978–997. [CrossRef]
43. Galambos, T.V.; Ellingwood, B.; MacGregor, J.G.; Cornell, C.A. Probability based load criteria: Assessment of current design practice. *J. Struct. Div.* **1982**, *108*, 959–977. [CrossRef]
44. DeBock, D.J.; Liel, A.B.; Harris, J.R.; Ellingwood, B.R.; Torrents, J.M. Reliability-based design snow loads. I: Site-specific probability models for ground snow loads. *J. Struct. Eng.* **2017**, *143*, 04017046. [CrossRef]
45. Liel, A.B.; DeBock, D.J.; Harris, J.R.; Ellingwood, B.R.; Torrents, J.M. Reliability-based design snow loads. II: Reliability assessment and mapping procedures. *J. Struct. Eng.* **2017**, *143*, 04017047. [CrossRef]
46. Bean, B.; Maguire, M.; Sun, Y.; Wagstaff, J.; Al-Rubaye, S.A.; Wheeler, J.; Jarman, S.; Rogers, M. *The 2020 National Snow Load Study*; Technical Report 276; Utah State University Department of Mathematics and Statistics: Logan, UT, USA, 2021. [CrossRef]
47. Bean, B.; Maguire, M.; Sun, Y. The Utah snow load study. *Civ. Environ. Eng. Fac. Publ.* **2018**, *3589*. Available online: https://digitalcommons.usu.edu/cee_facpub/3589 (accessed on 7 November 2022).
48. Ellingwood, B.; Redfield, R. Ground snow loads for structural design. *J. Struct. Eng.* **1983**, *109*, 950–964. [CrossRef]
49. Mo, H.; Dai, L.; Fan, F.; Che, T.; Hong, H. Extreme snow hazard and ground snow load for China. *Nat. Hazards* **2016**, *84*, 2095–2120. [CrossRef]
50. Mo, H.M.; Ye, W.; Hong, H.P. Estimating and mapping snow hazard based on at-site analysis and regional approaches. *Nat. Hazards* **2022**, *111*, 2459–2485. [CrossRef]
51. Structural Engineers Association of Colorado. *Colorado Ground Snow Loads*; SEAC: Colorado, CO, USA, 2007.
52. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]

53. Diciccio, T.J.; Romano, J.P. A review of bootstrap confidence intervals. *J. R. Stat. Soc. Ser. B (Methodol.)* **1988**, *50*, 338–354. Available online: https://www.jstor.org/stable/2345699 (accessed on 7 November 2022). [CrossRef]
54. Davison, A.C.; Hinkley, D.V. *Bootstrap Methods and Their Application*; Cambridge University Press: Cambridge, UK, 1997. Available online: https://www.cambridge.org/core/books/bootstrap-methods-and-their-application/ED2FD043579F27952363566DC09CBD6A (accessed on 7 November 2022).
55. Hall, P. *The Bootstrap and Edgeworth Expansion*; Springer Science and Business Media: New York, NY, USA, 1992. Available online: https://people.eecs.berkeley.edu/~jordan/sail/readings/edgeworth.pdf (accessed on 7 November 2022).
56. Kim, J.H. Forecasting autoregressive time series with bias-corrected parameter estimators. *Int. J. Forecast.* **2003**, *19*, 493–502. [CrossRef]
57. Franco, G.C.; Reisen, V.A. Bootstrap approaches and confidence intervals for stationary and non-stationary long-range dependence processes. *Phys. A Stat. Mech. Its Appl.* **2007**, *375*, 546–562. [CrossRef]
58. Engsted, T.; Pedersen, T. Bias-correction in vector autoregressive models: A simulation study. *Econometrics* **2014**, *2*, 45–71. [CrossRef]
59. Palm, B.G.; Bayer, F.M. Bootstrap-based inferential improvements in beta autoregressive moving average model. *Commun. Stat.-Comput.* **2017**, *47*, 977–996. [CrossRef]
60. De Vos, I.; Everaert, G.; Ruyssen, I. Bootstrap-based bias correction and inference for dynamic panels with fixed effects. *Stata J.* **2015**, *15*, 986–1018. [CrossRef]
61. Everaert, G.; Pozzi, L. Bootstrap-based bias correction for dynamic panels. *J. Econ. Dyn. Control* **2007**, *31*, 1160–1184. [CrossRef]
62. Kim, J. Bias-corrected bootstrap inference for regression models with autocorrelated errors. *Econ. Bull.* **2005**, *3*, 1–8. Available online: http://www.accessecon.com/pubs/EB/2005/Volume3/EB-05C20017A.pdf (accessed on 7 November 2022).
63. Ferrari, S.L.; Cribari-Neto, F. On bootstrap and analytical bias corrections. *Econ. Lett.* **1998**, *58*, 7–15. [CrossRef]
64. Menne, M.; Durre, I.; Korzeniewski, B.; Vose, R.; Gleason, B.; Houston, T. *Global Historical Climatology Network—Daily (GHCN-Daily)*; Version 3.26; NOAA: Silver Spring, MD, USA, 2012. [CrossRef]
65. Natural Resources Conservation Service. *Snow Telemetry (SNOTEL) and Snow Course Data and Products*; NRCS: Washington, DC, USA, 2020. Available online: https://www.wcc.nrcs.usda.gov/snow/ (accessed on 7 November 2022).
66. Maine Geological Survey. *Maine Snow Survey Data*; Maine Geological Survey: Augusta, ME, USA, 2020. Available online: https://mgs-maine.opendata.arcgis.com/datasets/maine-snow-survey-data (accessed on 7 November 2022).
67. Northeast Regional Climate Center. *New York Snow Survey Data*; Data Obtained through Personal Correspondence with Northeast Regional Climate Center; NRCC-Cornell University: Ithaca, NY, USA, 2020. Available online: http://www.nrcc.cornell.edu/ (accessed on 7 November 2022).
68. Wheeler, J. Supplementary files for "Creating a Universal Depth-To-Load Conversion Technique for the Conterminous United States using Random Forests". *Ann. Arbor.* **2021**, *1001*, 48109. Available online: https://digitalcommons.usu.edu/all_datasets/177/ (accessed on 7 November 2022).
69. Zhang, H.; Nettleton, D.; Zhu, Z. Regression-enhanced random forests. In Proceedings of the JSM Proceedings, American Statistical Association, Alexandria, VA, USA, 23 April 2019; Section on Statistical Learning and Data Science; pp. 636–647. Available online: https://arxiv.org/pdf/1904.10416.pdf (accessed on 7 November 2022).