Bootstrap methods for bias correcting probability distribution parameters characterizing extreme snow accumulations

Kenneth Pomeyie (kenneth.pomeyie@usu.edu)
Utah State University

Brennan Bean
Utah State University

Research Article

Keywords: Bias Correction, Extreme value analysis, Snow Water Equivalent, Bootstrap, Simulation study

Posted Date: November 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3586261/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Bootstrap methods for bias correcting probability distribution parameters characterizing extreme snow accumulations

Kenneth Pomeyie\(^1\)* and Brennan Bean, PhD\(^1\)

\(^1\)Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, 84322, Utah, USA.

*Corresponding author(s). E-mail(s): kenneth.pomeyie@usu.edu; Contributing authors: brennan.bean@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 (\(\sigma\)) 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 is able to 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, Bootstrap, Simulation study
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, snow water equivalent (SWE) Al-Rubaye et al (2022). However, direct observations of SWE are currently limited across 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).

Another way to overcome data scarcity is to estimate 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 SWE at locations only recording snow depth (see Jonas et al (2009), Sturm et al (2010), and McCreight and Small (2014) 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).

The variance underestimation issue of modeled SWE is inevitable mathematically. To demonstrate this, let \( Y \) be a random variable that represents SWE and let \( X \) represent a \( n \times p \) covariate matrix of other climate and snow information that could be used to estimate SWE. We can formulate the regression model as follows:

\[
Y = f(X) + \epsilon
\]

where \( \epsilon \) is the error term with observations that are assumed to be independent and identically distributed (iid) realizations from a normal distribution with a zero mean and a constant variance \( (\sigma^2) \). A standard regression analysis tries to estimate the conditional mean \( E(Y|X) \) by minimizing the residual sum of squares for the available data. By the law of total expectation, as presented in Ross (2010), the mean of \( Y \) is \( E(Y) = E(E(Y|X)) \). On the other hand, the variance of \( Y \) can be represented as:

\[
\text{Var}(Y) = \text{Var}(E(Y|X)) + E(\text{Var}(Y|X))
\]

\[
= \text{Var}(f(X)) + E(\sigma^2)
\]
Table 1: Examining the impact of standard deviation variations in a log-normal distribution on the 98th and 99th percentiles. The mean parameter ($\mu$) of the distribution is fixed at zero, while $\sigma = 1$ is used as a reference point for comparison.

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>Extreme Event</th>
<th>Relative Increase (%)</th>
<th>98th percentile</th>
<th>Relative Increase (%)</th>
<th>99th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.8</td>
<td>10.24</td>
<td>124</td>
<td>12.9</td>
<td>26</td>
</tr>
<tr>
<td>1.1</td>
<td>9.57</td>
<td>12.9</td>
<td>9.57</td>
<td>21.1</td>
<td>51.2</td>
</tr>
<tr>
<td>1.2</td>
<td>11.8</td>
<td>16.3</td>
<td>11.8</td>
<td>31.2</td>
<td>59.1</td>
</tr>
<tr>
<td>1.3</td>
<td>14.4</td>
<td>20.6</td>
<td>14.4</td>
<td>51.2</td>
<td>100.1</td>
</tr>
</tbody>
</table>

\[ = Var\left(f(\hat{X})\right) + \sigma^2 \]  

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

\[ Var(\hat{Y}) = Var\left(f(\hat{X})\right) < Var(Y) \]  

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 load designs. To demonstrate this challenge, we consider Table 1, which showcases the impact of varying the standard deviation ($\sigma$) 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 annual maximum SWE.

In this paper, we address the under-estimation of variance in SWE estimation models by introducing a residual bootstrap bias correction technique. We demonstrate in Section 4 that our approach is easier to generalize and scale as compared to existing approaches. This paper also provides comparisons of station-level distribution estimates with and without our bootstrap correction. In addition to this, we compare the station-level distribution of climate reanalysis data with and without the use of the Empirical Cumulative Distribution Function (ECDF) correction method, as previously introduced by Cho and Jacobs (2020).

This paper proceeds as follows: We first present the background literature on bias correction. We then detail the process of our proposed residual bootstrap technique for correcting bias in the distribution scale and demonstrate its effectiveness through a set of simulations and real-world data analysis. The paper concludes with a discussion of the implications of our findings and potential areas of future research.
2 Background - 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 (Zhang and Lu, 2012).

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., Fig. 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) incorporated an approach based on residual bootstrap to reduce bias for ensemble methods. The proposed method led to substantial reductions in both bias and improvements in predictive accuracy. Similarly, Breiman (1999) proposed an adaptive bagging technique to reduce prediction bias in ensemble models. Despite its strengths, this technique has a potential limitation of an increased variance that can offset the decrease in bias pointed out by Breiman (1999). 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) 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 (Zhang and Lu, 2012; Song, 2015; Ghosal and Hooker, 2020) 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).

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 the 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 $p \in [0,1]$. Mathematically, the approach is expressed as:

$$x_{cor} = t(x_{est}) = F_{obs}^{-1}(F_{est}(x_{est}))$$ (3)

An example of this approach is demonstrated in Cho and Jacobs (2020), where a 4km SWE reanalysis product developed by the University of Arizona (Broxton et al, 2016), hereafter referred to as UA SWE, is bias-corrected using the 1km SNOW Data Assimilation System (SNODAS) project (Barrett, 2003) developed by the National Weather Service’s (NWS) National Operational Hydrologic Remote Sensing
Figure 1: A scatterplot showcasing the systematic bias associated with random forest model at the Grassy Lake weather station in Wyoming. The estimated load: represents the estimated snow load predicted using a random forest model, and the actual load: represents the observed ground snow load.

Center (NOHRSC). The UA SWE data is created through the assimilation of numerous ground-based measurements of SWE from the Snowpack telemetry (SNOTEL) and the NWS Cooperative Observer Program (COOP) network sites and 4km grid-ded 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 (Carroll et al., 2001).

In the Cho and Jacobs (2020) study, the bias correction process considers the annual maximum SNODAS SWE data (1km x 1km aggregated to 4km x 4km spatial grid) from October 2003 to May 2017 (14 snow years) and the annual maximum UA SWE data (4km x 4km 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:

$$ UA_i^* = \text{CDF}^{-1}_{\text{SNODAS}^+}(\text{CDF}_{\text{UA}^\mp}(UA_i)), \text{ if } UA_i \leq \max(\text{UA}^\mp) $$

(4)
where $UA_i^*$ and $UA_i$ represent the bias-corrected and original UA SWE respectively at year $i$, $CDF_{SNODAS \pm}$ and $CDF_{UA \pm}$ represent inverse CDF of SNODAS SWE and CDF of UA SWE for the overlapping years, and $\max(UA_\pm)$ given as the maximum UA SWE in the overlapping years.

If $UA_i$ at year $i$ exceeds $\max(UA_\pm)$, the transformation is applied as follows:

$$UA_i^* = \max(SNODAS_\pm) + \left[UA_i - \max(UA_\pm)\right] \cdot \frac{\sigma(SNODAS_\pm)}{\sigma(UA_\pm)}$$

(5)

where $\max(SNODAS_\pm)$ given as the maximum SNODAS SWE in the overlapping years, and $\sigma(SNODAS_\pm)$ and $\sigma(UA_\pm)$ 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 ([Cho and Jacobs, 2020; Wood et al, 2004; Boé et al, 2007]). 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 ([Ashfaq et al, 2010]). Furthermore, the probability mapping method has been extensively used in climate modeling studies to correct bias ([Haerter et al, 2011; Piani et al, 2009; Ines and Hansen, 2006]). 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) 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 is not available, as will be demonstrated in the following sections.

3 Background - Reliability Targeted Loads

Before the introduction of ASCE 7-22 (2022) 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), with more details given in Ellingwood et al (1982) and Galambos et al (1982). Although the initial calibration was correct, research by DeBock et al (2017) and Liel et al (2017) 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 failure thresholds defined in Chapter 1 of ASCE 7-22 (2022). 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) shows a workflow of the reliability-targeted load simulation process.

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

$$1.2D + 1.0S = \phi R.$$  \hspace{1cm} (6)

where $D$ represents dead load, $S$ represents roof snow load, and $R$ 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

$$S = G_l * G_r$$  \hspace{1cm} (7)

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). In ASCE 7-22 (2022), 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 (DeBock et al, 2017; Bean et al, 2018; Ellingwood and Redfield, 1983; Mo et al, 2016, 2022) and gamma distribution (SEAC, 2007) have been employed to fit $G_l$. The ultimate objective is to select a distribution that effectively describes the entire history of 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 annual maximum SWE. Section 4 demonstrates our attempt to mitigate the bias in the scale parameter estimates of the site-specific probability distributions associated with $G_l$. 
4 Methods

The appeal of the bootstrap method, originally introduced by Efron (1979), is in its wide applicability to complex data structures in both parametric and non-parametric problems, as pointed out by Diciccio and Romano (1988). Both Davison and Hinkley (1997) and Hall (1992) 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) 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 (Kim, 2003; Franco and Reisen, 2007; Engsted and Pedersen, 2014; Palm and Bayer, 2017), in dynamic panel models (De Vos et al., 2015; Everaert and Pozzi, 2007), and regression models (Kim, 2005) to rectify bias in various estimators. It has also been used in machine learning models to minimize bias in the predicted response variable (Hooker and Mentch, 2016) and in the maximum likelihood estimation (MLE) (Ferrari and Cribari-Neto, 1998). 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 prediction error (Everaert and Pozzi, 2007). 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 (Wheeler et al., 2022). In this paper, both standard deviation and scale will be represented using the symbol $\sigma$. 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 1.

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 $X$. In order to correct for the bias in the distribution parameter ($\sigma$) of $Y$, we construct multiple distribution parameters over $B$ bootstrap replicates using a residual bootstrap according to the following algorithm:

1. Train/fit $f(X)$ using available data.
2. Obtain residuals $\epsilon = Y - \hat{f}(X)$.
3. Fit the true $Y_i$ to a specified distribution and obtain the initial scale parameter of the distribution $\sigma^0$.
4. Letting $b$ represent one of $B$ bootstrap samples:
   * Obtain new responses by bootstrapping the residuals and adding to the predicted responses.
\[ Y^b = f(\tilde{X}) + \epsilon^b \]

where \( Y^b \) represents a simulated alternative to \( Y \) based on a bootstrap sample of residual terms.

- Use \( Y^b \) to estimate the parameters of the specified distribution \( (\sigma^b) \)

5. Find the percentile \( (p^*) \) of the estimates of \( \sigma^1, \ldots, \sigma^b \) that is closest to \( \sigma^0 \). The closest scale parameter is denoted as \( \sigma^* \).

In the context of spatial analysis where the response variable \( Y_i \) can be unobserved, the discussed algorithm utilizes \( f(\tilde{X}) \) and an estimate of \( \epsilon^b \) to construct various estimates of \( \sigma \) via the procedure outlined in the algorithm’s fourth step. Following this, a \( p^* \), deduced from observably proximate locations, is employed to ascertain an unbiased scale parameter value. Notably, the \( p^* \) derived from these observable nearby locations can be calculated as either a global or a regional average, providing flexibility in accommodating various spatial analysis contexts and data availability scenarios.

5 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 a 10,000 and 1,000 sample size, with the following form:

\[ Y = \beta_0 + \sum_{j=1}^{n} \beta_j X_j + \epsilon \quad (8) \]

For each model, we partition the observations into a 70-30 training/test split. The first model is a 10-dimensional Gaussian model with a variance of 1.8 for each feature, and coefficients \( \beta = [0, 1, \ldots, 1] \). We generate the Gaussian model 200 times for each simulation, with the covariance between each pair of features ranging from 0 to 0.5. Additionally, each feature follows a normal distribution with a mean \( (\mu = 5) \) and standard deviation \( (\sigma = 5) \). The error term, \( \epsilon \), is drawn from a normal distribution with a mean of 0 and a standard deviation of two.

The second model is a 5-dimensional uniform model with a linear structure similar to the equation described by Equation 8. The coefficients for this model are set as follows: \( \beta = [-5, 4, 2.5, 5, 2, 1] \). We generate the uniform model 200 times for every 10,000 and 1,000 simulations, where each feature is uniformly distributed between -1 and 1 with equal probability. 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 follows varying distribution assumptions.
5.1 Experimental Setup

We train separate linear regression models for the two simulated training data sets. 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 $\sigma$. Note that we adjust $\sigma$ here because it is the parameter that is systematically underestimated in our application of interest. This is achieved by:

- Selecting the best percentile ($p^*$) from the sampling distribution of the $\sigma$ parameter that corrects for the bias in $\sigma$ for the training data.
- Generating the sampling distribution of $\mu$ and $\sigma$ using the test data set with a bootstrap sample size of 200.
- Using $p^*$, selecting the least biased $\sigma$ parameter along with its associated $\mu$ 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.

5.2 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 above one implies no bias, whereas a ratio below one signifies an underestimation and vice versa. The second quadrant (set of boxplots) of the figure reveals that “unadjusted” parameters consistently underestimate the value of $\sigma$ when the data generation sample size is 10,000. Conversely, the first quadrant shows that our bias-correction method effectively mitigates this bias in $\sigma$ (at $n = 10,000$), albeit at the expense of a minor increase in the variance of the ratios for $\mu$. This pattern — bias elimination in $\sigma$ and a slight increase in variance for $\mu$ — is also evident in the third and fourth quadrants, which are associated with a data generation sample size of 1,000. These results demonstrate that our technique can achieve some form of bias correction, regardless of the sample size.

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, 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 $\mu$ is estimated without bias — evidenced by a ratio value of one — regardless of whether bias correction is applied. Conversely, $\sigma$ 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 $\sigma$ up to one across different sample sizes. Regarding
Figure 2: Boxplots comparing the ratio of the actual normal distribution parameters (mean and SD) to parameters estimated using the linear model predictions. The data generation models: uniform and Gaussian models are generated at different sample sizes. Ratios are shown both before (unadjusted), and after (adjusted) the bias correction technique has been 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: 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>COV</th>
<th>Bias</th>
<th>COV</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>μ</td>
<td>σ</td>
<td>μ</td>
<td>σ</td>
</tr>
<tr>
<td>n = 10,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.01</td>
<td>0.01</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>Adjusted</td>
<td>0.01</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>n = 1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unadjusted</td>
<td>0.04</td>
<td>0.04</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>Adjusted</td>
<td>0.03</td>
<td>0.03</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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 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 3: Description of variables used in the model development phase as originally presented in Wheeler et al (2022). The data ranges from 1915-2020

<table>
<thead>
<tr>
<th>Climate Variables</th>
<th>Description</th>
<th>Units</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>Snow Depth</td>
<td>mm</td>
<td>$h_t$</td>
</tr>
<tr>
<td>SWE</td>
<td>Snow Water Equivalent</td>
<td>mm</td>
<td>$s$</td>
</tr>
<tr>
<td>MCMT</td>
<td>Mean Coldest Month Temperature</td>
<td>$^\circ$C</td>
<td>$T_c$</td>
</tr>
<tr>
<td>MWMT</td>
<td>Mean Warmest Month Temperature</td>
<td>$^\circ$C</td>
<td>$T_w$</td>
</tr>
<tr>
<td>TD</td>
<td>MWMT - MCMT</td>
<td>$^\circ$C</td>
<td>$T_d$</td>
</tr>
<tr>
<td>PPTWT</td>
<td>Sum of Winter Precipitation (Dec - Feb)</td>
<td></td>
<td>$P_t$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-Climate Variables</th>
<th>Description</th>
<th>Units</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2C</td>
<td>Distance to Coast</td>
<td>m</td>
<td>$D_c$</td>
</tr>
<tr>
<td>ELEV</td>
<td>Elevation</td>
<td>m</td>
<td>E</td>
</tr>
<tr>
<td>SMONTH</td>
<td>Month of Max Depth (Oct - 1, Jun - 9)</td>
<td></td>
<td>$M_s$</td>
</tr>
</tbody>
</table>

6 Application

Section 1 describes the inevitable underestimation of true SWE variance that comes when using estimates of 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 comes from Wheeler et al (2022), which proposed a RF model to estimate the annual maximum snow load (or equivalently, SWE) from the annual maximum snow depth. The data set includes annual maximum pairs of SWE/snow depth from the global historical climatological network (GHCND) (Menne et al, 2012), which includes both First Order Stations (FOS) and Snowpack telemetry (SNOTEL) stations. The consolidated dataset also includes Snow Course (SC) observations from the Natural Resources Conservation Service (NRCS, 2020). Note that the maximum snow depth and maximum SWE need not occur on the same day of the snow season. Additionally, the dataset also includes region-specific snow measurements from Maine (ME) (Maine Geological Survey, 2020), New York (NY) (NRCC, 2020), and California (CDWR, 2020). Please see Wheeler (2021) for access to this dataset. Table 3 describes all the variables used in the model development phase, including ancillary variables (such as temperature and distance to the coast) used in the RF model building phase. The data spans from the period 1915 - 2020, with 45,879 observations from 2,473 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.

6.1 Depth-to-load Conversion model

Using these data, we build a RF depth-to-load conversion model similar to the model built by Wheeler et al (2022). 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 9, where a prediction of snow water content
Table 4: Number of qualified \((n \geq 20)\) weather stations within each weather station network.

<table>
<thead>
<tr>
<th>Network</th>
<th>Number of weather stations</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOS</td>
<td>46</td>
</tr>
<tr>
<td>ME</td>
<td>23</td>
</tr>
<tr>
<td>NY</td>
<td>122</td>
</tr>
<tr>
<td>SC</td>
<td>144</td>
</tr>
<tr>
<td>SNOTEL</td>
<td>74</td>
</tr>
</tbody>
</table>

The ratio \(\eta_{\text{model}}\) is multiplied by snow depth \(h_t\) to estimate SWE.

\[
\text{SWE}_t = h_t \times \eta_{\text{model}}(h, T_c, T_w, T_d, P_t, D_c, E, M_s) + \epsilon_t
\]  

(9)

The response variable for the snow density model is a ratio of the SWE to snow depth \(\left(\frac{\text{SWE}}{h_t}\right)\), 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 SWE, as described in Sturm et al (2010). 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 then estimate the parameters of the log-normal distribution once using the direct measurements of SWE, and again using the estimated measurements of SWE using the three previously described machine learning models. For each distribution, we estimate the distribution parameters using maximum likelihood estimation. Figure 4 displays a boxplot of the estimated parameter ratios, specifically \(\mu^*\) and \(\sigma^*\). Here, \(\mu^*\) and \(\sigma^*\) correspond to the distribution parameters for the observed loads, while \(\mu\) and \(\sigma\) denote those for the estimated loads, according to the depth-to-load conversion method utilized in the RF, SVR, and GBM models. The scale/variance parameter (i.e., \(\sigma\)) tends to be consistently undervalued for the FOS, ME, NY, and SNOTEL weather station networks, as illustrated in the “unadjusted” section of Figure 4. Considering that \(\sigma\) 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.

6.2 Bias Correction in Snow Load Distribution

Figure 4 shows the results of our proposed bias correction method, described in Section 4, when applied to the different machine learning models. To achieve this, we first train conversion models to predict 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 SWE are used to obtain estimates of \(p^*\). We recognize that even with 20 observations, the parameter
Figure 3: Boxplots comparing the estimated log-normal distribution parameters (µ and σ) using the RF-estimated versus actual snow loads 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 fit to each weather station with the sample size of each network provided in Table 4.

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 comprises 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 corresponds with the σ estimates obtained from direct measurements of 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 4 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 ($p^*_{\text{avg}}$) which is 0.67, an average of the $p^*$’s. The parameters of the log-normal distribution displayed in Figure 4 show the before and after 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’s across the different stations cohort for each machine learning model are found to be underestimated. In contrast, the mean parameter values (µ) typically exhibit less bias.
Figure 4: Boxplot comparing the estimated log-normal distribution parameters ($\mu$ and $\sigma$) using the estimates versus actual loads when biased is corrected. Ratios are shown both before (unadjusted), and after (adjusted) the bias correction technique has been 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 fit to each weather station with the sample size of each station cohort provided in Table 4. The depth-to-load conversion models include the gradient boosting machine (GBM), RF, and support vector regression (SVR).

However, there is a notable exception for the mean values ($\mu$) in the case of the GBM and RF models at the first-order station cohort. The “adjusted” portion of Figure 4 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 $\sigma$ 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 $\sigma$ or $\mu$ 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 facilitated full bias correction, in contrast to relatively less effective models such as the GBM and RF, which did not fully achieve the desired bias correction. This highlights the important role of model choice in the successful implementation of our bias correction approach.

We also evaluated our bootstrap bias correction methodology using the GEV distribution across diverse station cohorts for each machine learning model. The process for bias correction remained identical, the only difference being the utilization of the GEV distribution for distribution estimation instead of the log-normal distribution.
Figure 5 shows the “unadjusted” and “adjusted” parameters when the different machine learning models are used. Just like Figure 4, 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 4. Nevertheless, following the bias correction, we see that the bias in the scale parameter ($\sigma$) 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 depicts the evaluation of the shape parameter ($\xi$) of the GEV distribution before and after bias correction using a relative change defined as

$$\text{Relative change} \left(\xi_d, \xi_e\right) = \frac{\xi_e - \xi_d}{|\xi_d|}$$

where $\xi_d$ and $\xi_e$ are the shape parameters from the distribution of the direct and estimated SWE, respectively. Figure 6 illustrates a comparison between unadjusted and adjusted $\xi$ parameters across various weather stations and machine learning models. A key observation from the figure is the inadequate description of the $\xi$ parameter both before and after scale parameter adjustment, indicated by the majority of ratios deviating from the reference line. One plausible explanation for this could be the insufficient number of observations used for distribution fitting. As emphasized by Mo et al (2022), accurately fitting all three parameters of the GEV can be challenging when dealing with small sample sizes. Also, comparing boxplots for the unadjusted and adjusted shows similar patterns. This similarity suggests that the proposed bootstrap approach, which primarily focused on the scale parameter, had no noticeable effect on the $\xi$ parameter. This outcome aligns with our expectations since the approach specifically targeted the scale parameter rather than the $\xi$ parameter.

### 6.3 RTL Results after Bias Correction

In this section, we assess the impact of our bias correction on the RTLs calculated in Bean et al (2021). RTLs, as explained in Section 3, 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 in Bean et al (2021), 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 $B = 500$ were repeated for 6,727 Tier 1 and 510 Tier 2 stations described in Bean et al (2021). Tier 3 stations from that report were excluded from this analysis due to their insufficient number of non-zero observations, rendering 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 was only applied to indirect measurements of snow load, and direct measurements of snow load remained identical in each iteration of the bootstrap sampling.

Figure 7 visually represents the percentage change in RTLs that results when using the bias-corrected probability distribution parameters. When focusing on the interquartile range, 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 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 snow depth and SWE in order to estimate the bias-correction percentile.
Figure 6: Boxplot comparing the estimated shape parameter of the GEV distribution using estimated versus measured loads when bias is corrected. Ratios are shown both before (unadjusted), and after (adjusted) the bias correction technique has been 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 fit to each weather station with the sample size of each station cohort provided in Table 4. The depth-to-load conversion models include the gradient boosting machine (GBM), random forest (RF), and support vector regression (SVR).

precisely. This would be challenging as most weather stations with co-located depth and SWE measurements are predominantly situated in mountainous regions.

7 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 5. Zhang et al (2019) 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 8 illustrates a single fit of the RF model applied to the linear problem formulated in the Gaussian version of Equation 8. 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 9 shows that despite the attempt to correct the bias in the scale parameter ($\sigma$), 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.

Additionally, Figure 4 highlights another limitation of the bootstrap method. When the variance/scale parameter ($\sigma$) is already unbiased, such as is the case with the snow course 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) has its own limitations. As detailed in Section 2, this alternative bias corrects UA SWE using SNODAS SWE through CDF matching. The impact of this alternative correction is depicted in Figure 10. 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 10 illustrates that the unadjusted parameters result in some systematic underestimation of $\sigma$ for the FOS, ME,
8 Conclusion

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 data sets. 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 SWE can still yield unbiased scale parameters when SWE is estimated through a regression-based approach.
Figure 9: 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 has been employed. A ratio closer to one means that the parameter is less biased, while ratios below one indicate an underestimate and vice versa.

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.

Declarations

- **Conflict of interest/Competing interests** The authors declare no competing interests
- **Authors’ contributions** KP wrote the main manuscript text and prepared figures. BB supervised research and reviewed the manuscript.
Figure 10: Boxplot comparing the estimated log-normal distribution parameters ($\mu$ and $\sigma$) 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.

References


22


CDWR (2020) California snow survey data. Data obtained through personal correspondence with organization


23


NRCC (2020) New york snow survey data. URL http://www.nrcc.cornell.edu/, data obtained through personal correspondence with northeast regional climate center

NRCS (2020) Snow telemetry (SNOTEL) and snow course data and products. URL https://www.wcc.nrcs.usda.gov/snow/


26