Journal of Data Science 21 (2), 368–390
April 2023
DOI: 10.6339/23-JDS1086
Statistical Data Science

Comparing Extreme Value Estimation Techniques for Short-Term Snow Accumulations

Kenneth Pomeyie[1,*], Brennan Bean[1], and Yan Sun[1]

[1]*Department of Mathematics and Statistics, Utah State University, 3900 Old Main Hill, Logan, UT, 84322, USA*

Abstract

The potential weight of accumulated snow on the roof of a structure has long been an important consideration in structure design. However, the historical approach of modeling the weight of snow on structures is incompatible for structures with surfaces and geometry where snow is expected to slide off of the structure, such as standalone solar panels. This paper proposes a "storm-level" adaptation of previous structure-related snow studies that is designed to estimate short-term, rather than season-long, accumulations of the snow water equivalent or snow load. One key development associated with this paper includes a climate-driven random forests model to impute missing snow water equivalent values at stations that measure only snow depth in order to produce continuous snow load records. Additionally, the paper compares six different approaches of extreme value estimation on short-term snow accumulations. The results of this study indicate that, when considering the 50-year mean recurrence interval (MRI) for short-term snow accumulations across different weather station types, the traditional block maxima approach, the mean-adjusted quantile method with a gamma distribution approach, and the peak over threshold Bayesian approach tend to most often provide MRI estimates near the median of all six approaches considered in this study. Further, this paper also shows, via bootstrap simulation, that the peak over threshold extreme value estimation using automatic threshold selection approaches tend to have higher variance compared to the other approaches considered. The results suggest that there is no one-size-fits-all option for extreme value estimation of short-term snow accumulations, but highlights the potential value from integrating multiple extreme value estimation approaches.

Keywords *extreme value theory; generalized extreme value distribution; mean recurrence interval; peak over threshold; snow water equivalent*

1 Introduction

In many parts of the United States (U.S.), buildings must be designed to withstand harsh weather conditions. One environmental hazard of particular interest in mountainous and northern states is the weight of accumulated snow on a structure, also called the snow load. This quantity is directly correlated with the water content of the accumulated snow, often called the snow water equivalent (SWE). Historically, building design standards have been tied to x-year recurrence intervals that depend upon the importance of the structure. For most structures, the "design" snow load has been associated with a 50-year mean recurrence interval (MRI). However, recent

*Corresponding author. Email: kenneth.pomeyie@usu.edu.

Received August 1, 2022; Accepted January 17, 2023

engineering codes have begun to move away from the MRI-based approach in favor of "reliability-targeted" load approach (Bean et al., 2021). Nevertheless, extreme MRI snow loads still play a prominent role in engineering design and standards.

Traditional structure design for snow loads assumes a strong correlation between the weight of snow on the ground and the weight of snow on a nearby roof. However, this assumption may not hold for structures like ground-mounted solar panels (GMSP) with material properties that encourage the regular shedding of snow. The shedding effect limits the amount of time that snow can remain on the panels, which differs substantially from traditional roofing systems that presume that accumulated snow stays in place for extended periods. This creates a pressing need to understand the dynamics of both short and long-term snow accumulations on structures with slick surfaces.

Addressing this need requires a change in the temporal scale of the snow measurement considered. While current design loads are informed by annual maximum ground snow loads, short-term design loads require the consideration of load accumulations at a weekly or daily scale. This paper introduces a "storm-level" adaption of the extreme value distribution that aims to evaluate x-day, rather than annual, accumulations of ground snow load. To do this, we mimic the shedding of the snow accumulation process on panels by defining and extracting snow accumulations over 1 to 6-day periods. The remaining sections of the paper proceed as follows. We first present the background of extreme value theory in Section 2, followed by the data collection and preparation process in Section 3. In addition, Section 3 describes a random forests (RF) model for imputing SWE at weather locations where only snow depth has been measured. In Section 4, we summarize the process of extracting single- and multiple-day snow accumulations, followed by an exploration of multiple approaches for estimating the 50-year MRI at more than 3,000 stations across the conterminous U.S. Section 6 concludes with a discussion of the implications of these results for GMSPs and highlight future research opportunities to better characterize short-term snow accumulations.

2 Background

Extreme value models focus on the tail behavior of a random process, which is necessary for extrapolating future extreme events given a limited number of historical observations. The two widely used methods for analyzing extremes are generally referred to as block maxima (BM) and peak over threshold (POT). The BM method was described by Fisher and Tippett (1928), and later proven by Gnedenko (1943), forms the Fisher–Tippett–Gnedenko theorem. The theorem states that the maximum observations collected from blocks of observations in non-overlapping periods converges to a generalized extreme value (GEV) distribution under the assumption that the maximum values from each period are independent and identically distributed. The distribution combines the two-parameter Gumbel, Fréchet, and Weibull distributions into a single family of distributions bound together by a third shape parameter ξ. The GEV cumulative distribution function (CDF) is represented mathematically as:

$$GEV(x;\mu,\sigma,\xi) = \begin{cases} \exp[-(1+\xi(\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}] & \xi \neq 0, \\ \exp[-\exp(-(\frac{x-\mu}{\sigma}))] & \xi = 0. \end{cases} \tag{1}$$

The m-year MRI or return level is defined as the event whose magnitude is expected to be exceeded, on average, once every m-years. This value is derived from the probability distribution

describing the extreme events as:

$$1 - Pr(X \leqslant r_m) = \frac{1}{m}, \text{ where } r_m \text{ is m-year return level.} \tag{2}$$

For example, a 50-year MRI is a value for which the area under the curve to the right of the value is $1/50 = 0.02$. Replacing with the left-hand-side of (1) with (2) gives:

$$1 - \exp[-(1 + \hat{\xi}(\frac{\hat{r}_m - \hat{\mu}}{\hat{\sigma}})^{-\frac{1}{\hat{\xi}}}] = \frac{1}{m}$$
$$\exp[-(1 + \hat{\xi}(\frac{\hat{r}_m - \hat{\mu}}{\hat{\sigma}})^{-\frac{1}{\hat{\xi}}}] = 1 - \frac{1}{m}.$$

Solving for $\hat{r}_m$ gives an estimate of the m-year return level as:

$$\hat{r}_m = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}[1 - \{-\log(1-m)\}^{-\hat{\xi}}] & \xi \neq 0, \\ \hat{\mu} - \hat{\sigma}\log\{-\log(1-m)\} & \xi \to 0. \end{cases} \tag{3}$$

Alternatively, the slightly more modern POT method focuses on extreme events separated from the rest of the data, regardless of the time interval in which the event occurred. Pickands III (1975) and Balkema and De Haan (1974) state that for a properly selected threshold, the POT method is approximated by the generalized pareto distribution (GPD) represented mathematically as:

$$P(X > x | X > u) = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\sigma}} \tag{4}$$

then,

$$P(X > x) = \zeta_u\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\sigma}} \tag{5}$$

where ζ_u is the probability of exceeding a user-defined threshold u. The r_k return level of the POT method which is exceeded once every k observations is computed as:

$$\zeta_u\left[1 + \xi\left(\frac{x_k - u}{\sigma}\right)\right]^{-\frac{1}{\sigma}} = \frac{1}{k} \tag{6}$$

Rearranging Equation 6 yields:

$$\hat{r}_k = \begin{cases} \hat{u} + \frac{\hat{\sigma}}{\hat{\xi}}[\left(k n_x \hat{\zeta_u}\right)^{\xi} - 1] & , \xi \neq 0, \\ \hat{u} + \hat{\sigma}[\log\left(k n_x \hat{\zeta_u}\right)^{\xi}] & , \xi \to 0. \end{cases} \tag{7}$$

For an N-year return level, where n_x represents the number of observations per year, then $k = N * n_x$. Thus, the N-year return level is:

$$\hat{r}_N = \begin{cases} \hat{u} + \frac{\hat{\sigma}}{\hat{\xi}}[\left(N n_x \hat{\zeta_u}\right)^{\xi} - 1] & , \xi \neq 0, \\ \hat{u} + \hat{\sigma}[\log\left(N n_x \hat{\zeta_u}\right)^{\xi}] & , \xi \to 0. \end{cases} \tag{8}$$

Bommier (2014) states that ζ_u can follow a poisson distribution. Hence ζ_u is estimated as:

$$\hat{\zeta_u} = \frac{\hat{\lambda}}{n_x}$$

Reformulating Equation 8 in terms of λ, the m-year return level is restated as:

$$\hat{r}_N = \begin{cases} \hat{u} + \frac{\hat{\sigma}}{\hat{\xi}}[\left(N\hat{\lambda}\right)^{\xi} - 1] & \xi \neq 0, \\ \hat{u} + \hat{\sigma}[log\left(N\hat{\lambda}\right)^{\xi}] & \xi \to 0. \end{cases} \tag{9}$$

There is no consensus in the current extreme value theory literature as to whether POT or BM should be generally preferred. Bücher and Zhou (2021) argues that POT is preferable for quantile estimation, while BM is preferable for return level estimation. Ferreira and de Haan (2015) states that BM method is efficient when the observations are not exactly independent and identically distributed. On the other hand, the POT method is more flexible in applications where changing the block size is difficult (Ferreira and de Haan, 2015). The arbitrariness of selecting threshold(s) for the POT method is one of the key difficulties associated with the method. If the selected threshold is too low, the retained values will fail to give enough emphasis to the distribution tail behavior. On the other hand, a high threshold increases the variance of the estimators due to smaller sample sizes. For this reason, it is vital to set a threshold that finds a good balance between the bias and variance of the model. The literature proposes many diagnostic procedures for threshold choice. These procedures and their many variants can be roughly categorized into two forms: graphical methods and probabilistic-based or analytical methods. Coles (2001) outlines graphical methods that include the mean residual life (MRL) plot and parameter stability plot, among others. The MRL is a plot of the mean excess against a range of threshold values. Given the threshold stability property of the GPD, the plot should be linear above the suitable threshold for which the GPD model is valid. On the other hand, the parameter stability plot is based on the idea that if a threshold is suitable to approximate a GPD model, then the shape and modified scale parameters should remain stable above other thresholds. Figures 1 and 2 show an example of the MRL and parameter stability plot for a 1-day snow accumulation, measured via snow water equivalent (SWE) at the Pittsburgh International Airport, PA, weather station. In Figure 1, exceedances are approximately linear between threshold values of 17 mm and 25 mm. In Figure 2, the parameters appear to stabilize around a threshold of 23 mm, with subsequent values having a higher variance due to fewer exceedances. The combination of these two approaches suggests a threshold selection at or near 23 mm.

These graphical approaches to threshold selection have been criticized for their subjective nature (Coles, 2001; Scarrott and MacDonald, 2012; Yang et al., 2018). As an alternative, analytical methods have been proposed to automate the threshold selection process. Durrieu et al. (2015) proposed an automated threshold selection method based on a point-wise data-driven procedure to choose the threshold. Using a likelihood ratio test, thresholds are detected sequentially from observations residing in the right tail of the generalized pareto distribution. The test terminates once a tail is no longer detected at a statistically significant level. Solari et al. (2017) automates the threshold selection process based on the Anderson-Darling EDF statistic and goodness of fit test. As an alternative approach, Northrop et al. (2016) uses a Bayesian procedure to select the optimal threshold using a Bayesian implementation of leave-one-out cross-validation to compare the ability of the generalized pareto (GP) to correctly predict the density of withheld observations using a range of thresholds as compared to the densities obtained using a validation threshold. Zoglat et al. (2014), inspired by the work of Beirlant et al. (2005), developed an analytical method that selects the optimal threshold from a range of equally spaced thresholds that minimizes the square error between the simulated and the observed quantiles for a variety of probabilities. Zoglat et al. (2014) also used the likelihood ratio test to sequentially detect the

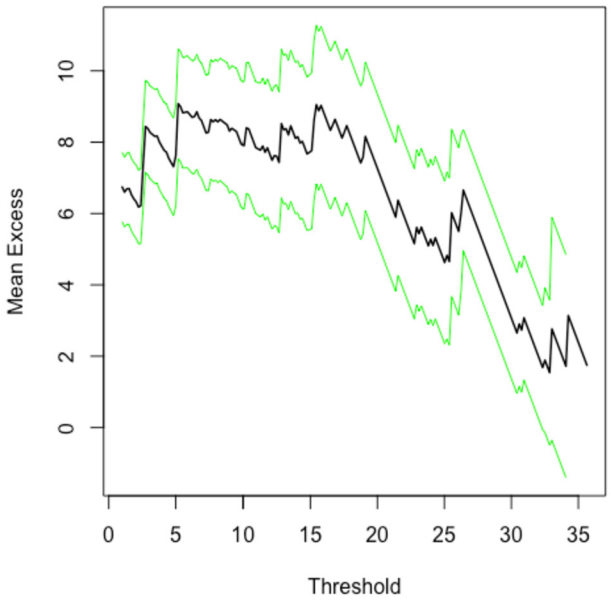


Figure 1: Mean residual life plot of 1-day snow accumulations at the Pittsburgh Intl. Airport weather station. The green lines represent the 95% confidence interval. Threshold values are in millimeters.

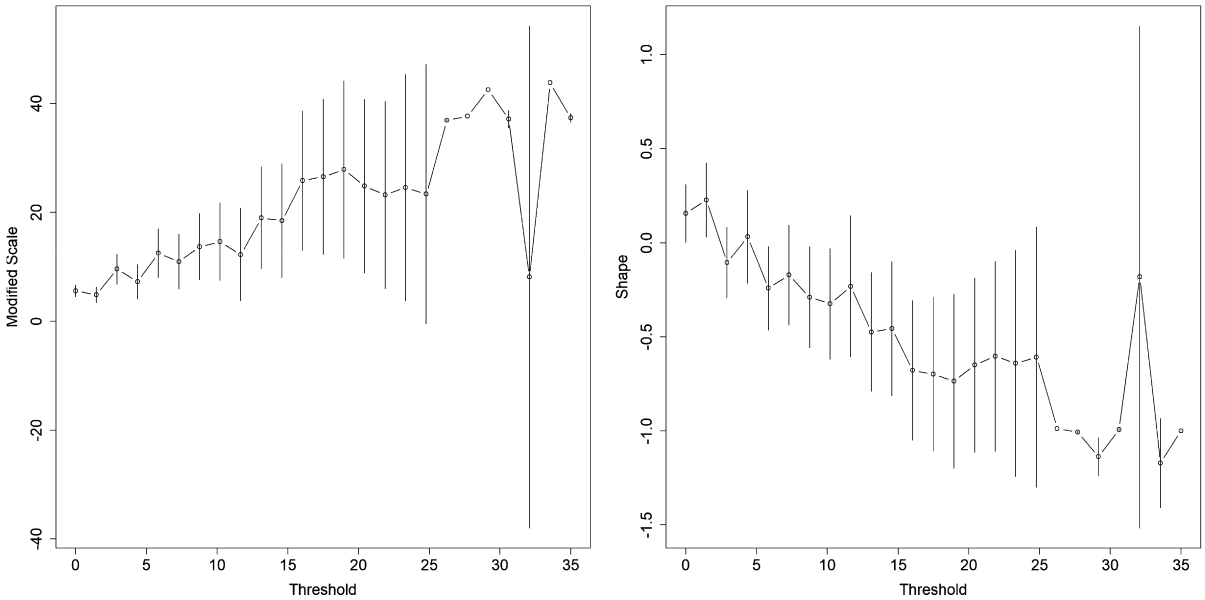


Figure 2: Parameter stability plot of 1-day snow accumulations at the Pittsburgh Intl. Airport weather station. The vertical line represents the 95% confidence interval. Threshold values are in millimeters.

optimal threshold from a range of potential thresholds. Rather than selecting an optimal threshold, Deidda (2010) used the concept of parameter threshold-invariance to develop a multiple threshold method that provides parameter estimates as median values of re-parameterizations

over a range of thresholds. This method was appropriate for regional data subjected to site-to-site variability. Curceac et al. (2020) developed an automated threshold selection method based on the threshold stability plot. This method fits a cubic smoothing spline and calculates the rate of change. The optimal threshold is the point on the plot where the modified scale and shape parameters reach a plateau. When compared to two previously discussed methods (Zoglat et al., 2014; Thompson et al., 2009), Curceac et al. (2020)'s method provided more robust parameter estimates for hydrological extremes. In this paper, we compare versions of the POT and mean-adjusted quantile (MAQ) method against the traditional block maxima (BM) method. For the POT version, we consider a zero threshold along with the threshold selection approaches of Northrop et al. (2016) and Durrieu et al. (2015). The mean-adjusted quantile method estimates extreme value using Equation 9. For the MAQ version, we consider the gamma distribution and the generalized extreme value distribution in the estimation process for extreme values.

3 Data Preparation and Estimation

All of the above-referenced approaches require direct measurements of the weight, or snow water equivalent, of snow in order to apply them to structure design. Unfortunately, direct measurements of the SWE are substantially more challenging to measure than other climate variables such as temperature, wind speed, and snow depth. As such, there are relatively few weather stations with consistent measurements of snow load to rely on in this analysis. The data sparsity problem is further exacerbated by data quality issues typical of any weather measurement, but especially problematic in an analysis focused on extreme values. This section describes the data collection process undertaken in this paper to screen out misrecorded measurements before analysis. We then separately present a random forests (RF) model that imputes missing values of snow load using the snow depth, snow duration, and other climate variables.

3.1 Data Collection

The daily snow depth (SNWD) and SWE measurements, both in millimeters, used in this analysis were obtained from the National Oceanic and Atmospheric Administration's (NOAA) Global Historical Climatology Network – Daily (GHCND) (Menne et al., 2012). This dataset includes measurements at National Weather Service (NWS) first-order stations (FOS), NWS co-operative observer (COOP) stations, and Snowpack Telemetry (SNOTEL) stations across the conterminous U.S. We examined snow seasons between 1858 and 2021, but not all stations were active for all 164 seasons. A snow season begins in November of the previous year and ends in May of the current year. Estimations of extreme x-day snow loads require long histories of continuous snow measurements. Because of this, stations were only retained if there was at least 50% measurement coverage (SWE or SNWD) of the snow season in at least ten different years. Using the measurement coverage filter, we obtained 6,245 stations (428 SNOTEL, 5,566 COOP, and 251 FOS) suitable for consideration in our analysis.

3.2 Quality Control

It is common for SNWD/SWE measurement pairs to express a physically impossible relationship within the historical record. These unrealistic measurements may cause serious misrepresentations of the distribution tail. Fortunately, many of these misreported values are already flagged by NOAA as provided in the GHCND data download Menne et al. (2012). All NOAA-flagged

Table 1: Percentage of missing values in snow information per station type. These percentages are computed for snow years.

	SNWD	SWE
FOS	13.3%	78.9%
COOP	9.9%	94.2%
SNOTEL	28.8%	3.7%

outliers were removed prior to analysis. While NOAA-flagged values significantly improve quality control, Bean et al. (2021) demonstrate that the NOAA quality controls do not flag many isolated and systemic data errors. Their efforts culminated in a manual list of misreported outlier observations flagged using interactive explorations of historical snow measurements facilitated by R's `plotly` package (Sievert, 2020). This paper removes all outlier values manually flagged in Bean et al. (2021) to ensure quality control. Additionally, stations were subjected to additional quality control procedures, including:

- SWE observations with negative values are clearly in error and removed from consideration.
- SWE observations with snow density above 0.8 (i.e. density of pre-glacial snow) are clearly outliers and are removed from considerations (Copland, 2022).
- In accordance with the precedent set by Tobiasson and Greatorex (1996), observations with snow density less than 0.05 are not used in our analysis.
- SWE observations below 10 millimeters (mm) that increase or decrease by a factor of 100 on a daily basis are removed from consideration. Such errors in snow measurement usually result from misplaced decimal points.
- SWE observations that show a 50% single-day increase when the previous day's observation is greater than 1000 mm are removed. Such a large daily increase exceeds the rainfall, let alone snowfall, records for most of the United States (Donegan, 2019) and are likely unrealistic.

3.3 Estimating SWE

The prevalence of direct measurements of SWE is highly related to geographic location and elevation. Table 1 shows that SWE data have a higher percentage of missing values than SNWD for FOS and COOP stations. This implies that some or most low-elevation stations do not measure the weight of snow directly. For this reason, estimating SWE from snow depth is necessary to characterize extreme loads for the entire country, and not just in the mountains of western states where the SNOTEL stations are located.

Estimating SWE from snow depth is difficult due to the complex relationship snow density shares with the local topography and climate (Sturm et al., 2010). The SWE/SNWD ratio at a point in time is referred to as the specific gravity of snow, which is directly related to the snow density (η_t). Numerous models have been developed in the literature as a general-purpose tool to estimate SWE directly, or snow density which is converted to SWE using snow depth. Jonas et al. (2009) developed a bulk density model using 12 different regression models to account for different altitude and snow season classes on a bi-weekly basis. Their model was trained on stations in the Swiss Alps, which likely have different climate and topography than most locations in the United States. Sturm et al. (2010) used a nonlinear ANCOVA model within a Bayesian framework to estimate monthly snow densities for snow climate classes. These conversion models were developed using only mountainous snowpack measurements and are not necessarily well

suited on a national scale or a daily basis. Wheeler et al. (2022) addresses the scaling issue using annual maximum snow depth/SWE data pairs as input into a RF model to capture complex nonlinear interactions between expected depth SWE ratios across the continental United States. However, this model is not intended for use on sub-seasonal data. On the other hand, McCreight and Small (2014) addresses the time step issue by developing a daily model using regression. This model predicts SWE using multi-equation models calibrated according to the month of the year and different climate classes. However, McCreight and Small (2014), like Sturm et al. (2010) trained their model using only information from mountainous snowpack, which tends to have greater snow densities than areas with mid-season snow melt. In light of this, we develop a random forests daily density model to impute SWE using SNWD and other climate information across the United States.

3.4 Snow Density RF model

This paper employs a RF model (Breiman, 2001) to approximate the complex relationship between snow density and other environmental factors. Given that SWE is highly correlated with SNWD, it is inefficient to model SWE directly, as the individual trees of the RF will be dominated by splits on the variable SNWD. Thus, modeling the specific gravity of snow allows us to focus on what factors influence the specific gravity of snow, which can then be multiplied by SNWD to recover SWE. Sturm et al. (2010) highlights that bulk density model errors are homoskedastic, while SWE errors are heteroskedastic.

The core dataset used to train and evaluate the snow density model came from conterminous United States' weather stations included in the GHCND Menne et al. (2012). This dataset includes only available SWE/SNWD pairs from November 1, 1980, to May 30, 2021, for a total of 41 snow seasons. The data is limited from the 1980 snow year because climate reanalysis information used in the RF model only extends as far back as 1979. The initial data collection resulted in nearly 3.5 million measurements in all conterminous forty-eight states. The snow measurements are supplemented with daily climate reanalysis data from gridMET CLIMATE source (Abatzoglou, 2011), which includes mean temperature, vapor pressure deficit, wind velocity, and solar radiation. The gridMET is a dataset of daily high-spatial resolution (4km) surface meteorological data covering the conterminous US. Table 2 contains the full list of variables used in model building, as well as their relative importance in the RF model predictions. A total of 2.6 million observations ($\approx 75\%$) are used to train the model with roughly 0.9 million ($\approx 25\%$) observations used for validation. Because the model is expected to impute SWE in different snow years and climates, observations are partitioned into training and test sets using snow years rather than randomly separating individual observations.

While RF permutation variable importance measures provide a sense of the individual influence of explanatory variables, it does not capture the interactive influence of the collection of variables. To explore such collective influence, we fit two different RF models to the training data. The two main hyper-parameters of the RF model – the minimum number of observations in a terminal node and the number of randomly selected variables considered for splitting at each mode – were tuned prior to conducting a 10-fold cross-validation. The first RF model, called the "Snow Density RF_1", is represented in Equation 10. This model incorporates information within the current accumulated snowstorm for which the SWE is required. The term accumulated snowstorm refers to snow that has accumulated on the ground from a single or several different snowstorm events. It is important to note that as snow builds up on the ground, its density is influenced both by climate variables and non-climate variables, resulting in different

Table 2: Description of predictor variables used in the RF regression model. The larger the value of "Importance", the greater the loss in predictive accuracy due to the loss of information from the associated variable in the RF model. The variable importance is computed on the storm-level RF on the full dataset (1980-2021).

Climate Variables				
Name	**Description**	**Units**	**Variables**	**Importance**
SRAD	Mean Solar Radiation	W/m^2	S_m	4660
TEMP	Mean Air Temperature	degK	T_m	4430
ROLL_TEMP	Rolling Average of TEMP	degK	T_r	2413
ROLL_VPD	Rolling Average of VPD	kPa	V_r	1541
VPD	Mean Vapor Pressure Deficit	kPa	V_m	1317
WIND	Mean Wind Velocity	m/s	W_m	927
Other Variables				
DAYS	Number of days from the current snow year		D_s	7155
SNWD	Snow Depth	mm	h	3413
GROUND	Number of days snow has been on ground		G	3391
D2C	Distance to Coast	m	D_c	2434
ELEV	Elevation of station	m	E_s	1995
ECO3	Level III Ecoregion		E_3	1239
ECO2	Level II Ecoregion		E_2	304
ECO1	Level I Ecoregion		E_1	209

accumulation and melting processes. Due to this reason, the Snow Density RF_1 model includes the information from the current day's estimation of the SWE and factors affecting the changing conditions of accumulated snow on the ground. With variables such as $ROLL\_TEMP$, $DAYS$, $GROUND$, and $ROLL\_VPD$, the RF can infer the changing conditions of accumulated snow, thereby improving its prediction capability. The second RF model, called the "Snow Density RF_2", is represented in Equation 11. The model is designed to incorporate general information on the day of the SWE estimation. That is, the RF model is trained based on the current climate and non-climate variables without accounting for changes in snow conditions within the accumulated snowstorm. Thus, both RF models take into account current-day of SWE estimation information when training the models. Besides current-day information, the Snow Density RF_1 also takes into consideration changes in accumulated snow from the beginning of the storm. This is the primary difference between the two models. Lastly, a classical regression model is trained to serve as a benchmark comparison using Equation 12. The variable definitions for all these equations are provided in Table 2.

$$\text{RF\_1} : \eta_t = f_1(h, T_m, T_r, V_m, V_r, S_m, W_m, D_c, E_s, E_1, E_2, E_3, D_s, G) + \epsilon_t \tag{10}$$

$$\text{RF\_2} : \eta_t = f_2(h, D_c, E_s, E_1, E_2, E_3, D_s, T_m, V_m, S_m, W_m) + \epsilon_t \tag{11}$$

$$\begin{aligned}\text{Regression} : \eta_t = {} & \beta_0 + \beta_1 h + \beta_2 T_m + \beta_3 T_r + \beta_4 V_m + \beta_5 V_r + \beta_6 S_m + \beta_7 W_m \\ & + \beta_8 D_c + \beta_9 E_s + \beta_{10} E_1 + \beta_{11} D_s + \beta_{12} G + \epsilon_t\end{aligned} \tag{12}$$

Table 3: Comparison of snow density estimation model on the test and train dataset. Error measures include the root mean square error (RMSE) and the mean absolute deviation (MAD).

	RF_1		RF_2		Regression	
	Train	Test	Train	Test	Train	Test
RMSE	0.0332	0.0436	0.0539	0.0663	0.0721	0.0714
MAD	0.0196	0.0285	0.0352	0.0459	0.0501	0.0502

Table 4: Percentage of missing values vs relative percent reduction per station type after SWE imputation. These percentages are computed for snow years.

	Missing SWE	Reduction in Missing SWE
FOS	13.1%	65.8%
COOP	35.3%	58.9%
SNOTEL	2.5%	1.2%

After training the three models, we used them to make predictions on the test data. A comparison of the model performance on the training and test data is provided in Table 3. It is apparent that the Snow Density RF_1 model has the smallest mean squared error (MSE) and mean absolute deviation (MAD) on the test data, suggesting that it is better compared to the regression model and the Snow Density RF_2 model. Hence, the Snow Density RF_1 model is chosen as the best model, and it is used as a general-purpose tool for imputing missing SWE values in this paper. The final model is re-trained using the full dataset. To impute the SWE, the predicted snow density ($\widehat{\eta}$) is multiplied by the snow depth. This implies that missing snow depth observations are propagated through the RF model. The percentage of missing SWE for snow seasons after imputation, along with the percent reduction in missing values relative to the data before imputation, are provided in Table 4.

4 Methodology

The data extraction process for single-day and multi-day change methods allows for consideration of accumulated snow loads over short periods of time rather than across an entire season. This allows us to investigate extreme value distributions under various snow-shedding scenarios, as might be expected on GMSPs or other slick, steep-pitched surfaces. The following subsections describe the methodology for extracting storm-level observations, along with the approach for estimating 50-year MRI short-term snow loads at weather stations across the conterminous U.S.

4.1 Daily Change Method

In order to analyze the storm-level accumulations of ground snow load, we analyze the positive sequential differences for a single or multi-day snow accumulations. This is different from analyzing the measured snowfall, as fallen snow may melt on contact with the ground and does not result in a structurally relevant accumulation of load. Further, note that the process only considered positive accumulations of snow, which differs from typical time series modeling that

considers all sequential changes in the measured variable of interest. As an example, let us assume a time series of daily ground snow load for a single snow year of length n as:

$$\{X_t\}_{t=1}^{n} = \{X_1, X_2, \ldots, X_n\}, \tag{13}$$

where $t = 1$ is the first day of the snow year and $t = n$ is the last day of the snow year.

To obtain the daily sequential changes, we take the first-order difference of the series shown in (13). In this paper, non-positive accumulations are assumed to imply the end of the snow load accumulation and are represented as zero. Mathematically, we define the length $n-1$ first-order difference of the series as:

$$\Delta X_{p_t} = \begin{cases} \Delta X_t & X_t > X_{t-1} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Single-day (D-1 or 1-day) Change method: The single-day change method assumes a perfect weather condition with enough sunlight that ensures snow melts off panels within 24 hours. Thus, the method retains positive observations from Equation 14 that represent daily snow on panels. Mathematically, we define the observations of the D-1 method from Equation 14 as:

$$\{\Delta X_{p_2}, \Delta X_{p_3}, \ldots, \Delta X_{p_k}\} \tag{15}$$

where $\Delta X_{p_i} > 0$ for $i = 1, 2, \ldots, k$, and k is the total number of positive changes.

Multi-day Change method: The single-day change method may be risky for use in engineering practice as it is not guaranteed that all snow will shed off of a slick surface (like a solar panel). There are many weather factors that may cause snow to "stick" between storms. As a result, we introduce the multi-day change method, which considers sequential consecutive positive differences in snowpack conditions over an arbitrary time. Given the sequential nature of the method, a D-2 method is made up of *up to* two consecutive sequential positive daily changes in a snow load. This implies that the D-2 day method includes both single day storms, as well as the sum of two consecutive single day storms when relevant. As an example, we define a consecutive ten-day series of daily changes from Equation 14 to describe the D-2 method. The following steps show how the multi-day change method captures D-2 observations.

$$\begin{gathered} \{\Delta X_{p_2}, \Delta X_{p_3}, \Delta X_{p_4}, \Delta X_{p_5}, \Delta X_{p_6}, 0, 0, \Delta X_{p_9}, 0, \Delta X_{p_{11}}\} \\ \{(\Delta X_{p_2} + \Delta X_{p_3}), (\Delta X_{p_4} + \Delta X_{p_5}), (\Delta X_{p_6}), (\Delta X_9), (\Delta X_{p_{11}})\} \end{gathered} \tag{16}$$

Equation 16 shows that sequential positive changes in snow accumulation are treated in blocks rather than using a moving window. Thus, if we have five days of positive consecutive differences, we will calculate the quantity $\Delta X_{p_2} + \Delta X_{p_3}$, $\Delta X_{p_4} + \Delta X_{p_5}$ and ΔX_{p_6}, but we will not calculate the quantity $\Delta X_{p_2} + \Delta X_{p_3}$. This limitation was imposed due to computational constraints, as the blocking approach allows for fast dataset segmentation. Additionally, sliding window approaches would make it so that some observations would be considered more than once in the distribution fitting, creating strong correlations between measurements that we are not currently equipped to address. Future studies should consider ways to overcome the computational and statistical barriers to incorporating moving window approaches.

This process can be extended to accumulation periods longer than two days. In this paper, the x-day method, where $x > 2$, extracts observations as a sum of x-consecutive positive daily sequential changes. The requirement of consecutive positive daily changes means that any reduction in accumulated snow within an x-day period will result in a reset of the x-day sequence.

Table 5: Percentage of storms eliminated by the snowfall indicator for Day-1,-2, and -3 loads. Snowfall indicator marks the start of the data extraction process.

	Day 1	Day 2	Day 3
FOS	55.2%	52.3%	51.3%
COOP	60.8%	57.2%	55.7%
SNOTEL	38.2%	33.3%	29.8%

This strategy assumes that any melting or reduction of snow will be associated with a "sliding" event on relevant structures, which will reset the snow accumulation totals to zero. For this reason, multiple-day periods are only counted if each day in the sequence has a positive snow accumulation. One future extension of this study would be to allow for slight reductions in the accumulated snow in the middle of an x-day period as long as the start-to-finish change in the snow load is larger than any of the individual one-day accumulations within the same period.

As part of our methodology, we distinguish legitimate increases in snow load caused by snowfall from random increases not caused by snowfall for the start of the individual extraction process. SWE, or snow load, is the water content of accumulated snow at a given point in time. It is impossible for the weight of accumulated snow to increase without snowfall feeding the snowpack. However, there are times when snow load (SWE) measurements show small increases simply due to a lack of measurement precision. These spurious increases in the snow load have the potential to infect our daily change method with an excess of small, positive daily changes. To avoid this, the variable snowfall is used to validate recorded increases in the snow load at a daily scale. This paper uses non-zero measurements of 24 hr snowfall as an indicator variable to trigger the start of a single and/or multiple-day snow accumulation event. In other words, a non-zero X_{p_t}, which is based on accumulated snow, is only retained if the weather station separately recorded a positive value for snowfall in any amount. This indicator variable is available for FOS and COOP stations but not SNOTEL stations. Due to a lack of information on snowfall data at SNOTEL stations, positive sequential changes above 3 mm are used to trigger a snow accumulation event. This was an arbitrarily selected threshold intended to capture small, yet legitimate, accumulations of snow while removing the majority of spurious snowpack increases. Table 5 shows the percentage of storms eliminated after accounting for snowfall. Since the trigger for a snow accumulation event at SNOTEL stations is fixed (at 3mm), the table shows a low storm removal rate when compared to FOS and COOP stations with a snowfall variable.

4.2 50-year Mean Recurrence Interval (MRI) Estimation

This paper defines an x-day "design" snow load as a 50-year MRI snow load event. We examine six different approaches using three different MRI estimation methods. The MRI estimation methods include the block maxima (BM), the mean-adjusted quantile (MAQ), and the peak over threshold (POT) method. Table 6 shows the six different approaches using the three different MRI estimation methods. The first method/approach, called the block maxima method, estimates design loads by fitting maximum annual x-day loads to a GEV distribution and extracting the 98th percentile. This method has long been acknowledged as data inefficient, as only one of the potentially many large storm events in a single snow season is retained. This is one of the primary arguments for the POT method to extreme value analysis, which considers all measured values over a user-defined threshold. One complication with using traditional POT to analyze

Table 6: Types of MRI estimation approaches explored in the analysis. Distributions include generalized pareto (GP), generalized extreme value (GEV), and gamma distribution.

MRI Estimation Approach	**Name**	**Distr.**
Block Maxima	BM	GEV
POT using zero threshold	ZERO	GP
POT using Northrop et al. (2016)'s threshold selection	BAYES	GP
POT using Durrieu et al. (2015)'s threshold selection	FREQ	GP
Mean Adjusted Quantile (MAQ): GEV	GEV	GEV
Mean Adjusted Quantile (MAQ): Gamma	GMA	Gamma

accumulated ground snow loads is the high correlations between daily snowpack measurements over the season. This is because the snow from the previous day is often fully present in the subsequent day's snow load value, creating large correlations in measurements of accumulated snow across time. Our method of looking at x-day changes is more conducive to the POT approach sequential changes in the snowpack across time are more likely to be independent.

The considered variations of the POT method are related to the processes used to select the threshold. For the second estimation approach, we consider a naive threshold of zero (ZERO). Here, we select a threshold of zero for the x-day load values to model the generalized pareto distribution. Evaluation of the sensitivity of the 50-year MRI events to the threshold selection allows us to highlight the importance of the threshold in modeling extremes. Consequently, the ZERO approach can be considered a benchmark for the third and fourth estimation approaches that employ an automated threshold selection algorithm. The third estimation approach (BAYES) automatically selects the optimal threshold using a Bayesian method proposed by Northrop et al. (2016). This methodology uses a Bayesian implementation of leave-one-out cross-validation to compare the predictive ability of generalized pareto (GP) inference based on different thresholds. In this framework, we validate training thresholds using no fewer than 50 threshold excesses when making inferences about a GP distribution as recommended by Jonathan and Ewans (2013). We consider 30 unique training thresholds due to our relatively small sample sizes with thresholds ranging from the 10th percentile to the largest observed value with at least 50 threshold excesses. The fourth estimation approach (FREQ) also uses an automated threshold selection process. The optimal threshold is selected using a frequentist method proposed by Durrieu et al. (2015). This method uses a point-wise data-driven procedure to select the threshold in two steps. The first step employs a sequence of likelihood ratio tests to identify a parametric fit to the GP distribution. Once a threshold is detected, the next step maximizes the penalized likelihood of the GP distribution to select an adaptive threshold that is less than the originally selected threshold. The penalization step is necessary to ensure that bias associated with the original threshold is reduced. For the different POT approaches, once the threshold is selected, the exceedances are fitted to the GP distribution. To estimate the 50-year design load, the 98th percentile is adjusted based on the average number of exceedances per snow year as demonstrated in Equation 9.

Lastly, we estimate the design loads using the mean adjusted quantile (MAQ) of a distribution. Like the POT method, the MAQ method fits distributions to daily x-day values, but the estimated quantile of the fitted distribution is adjusted based on the average accumulation events observed each year. The adjustment is necessary given that we need to estimate a 50-year

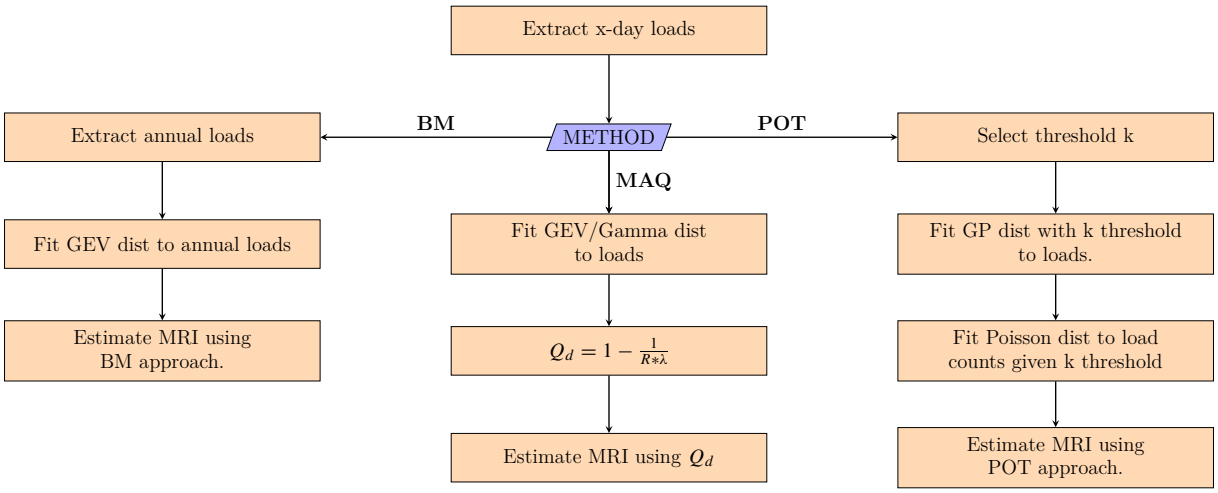


Figure 3: Workflow for estimating the 50-year mean recurrence interval (MRI) value from short-term loads. Distributions include generalized Pareto (GP), generalized extreme value (GEV), Gamma and Poisson. λ - mean of load count per snow season, R - return period (50 years).

event from daily data as shown in Equation 9. For the fifth and sixth estimation approaches (GEV and GMA), the mean adjusted quantile method is utilized, in which the GEV and gamma distributions are employed, respectively. These distributions are commonly used probability distributions for climate variables such as rainfall and snow. In Figure 3, we show a flowchart of the different methods employed in MRI estimation. We require a station with at least 30 observations to estimate the x-day design snow load, except for the POT Bayesian approach, which requires 200 observations. The 200 observations allow us to consider a wide range of potential thresholds while ensuring a minimally acceptable sample size in the tail of the distribution for inference.

5 Results

5.1 Short-Term Design Snow Load Estimates

In this study, we estimate the x-day design snow loads for several weather stations across the country. As discussed in the previous section, the design loads are estimated using six different estimation approaches from three different MRI estimation methods. Table 7 shows the number of qualified stations used in each MRI estimation approach. From the table, the BAYES approach has the lowest convergence rate despite requiring at least 200 observations for distribution fitting. Only stations with MRI values across all six approaches are included in the comparison.

For engineers, it is necessary to determine the maximum weight of snow that can collapse a structure, in this instance, ground-mounted solar panels. Historically, the design snow load has been associated with the 50-year MRI snow load event multiplied by a safety factor. Using the return level approach discussed in the background section, the 50-year event can be estimated by fitting a distribution to the data set and extrapolating from the tails of the distribution. The extrapolation process is necessary since the magnitude of the event value exceeds the amount of data that can be observed. Extrapolations are, however, difficult to validate directly since they

Table 7: Comparison of station convergence rate for each MRI estimation approach. See Table 6 for MRI estimation approaches names and descriptions.

		MAQ (GEV)	BM	POT (BAYES)	MAQ (GMA)	POT (FREQ)	POT (ZERO)
Day 1	Number of stations	3,124	3,124	1,696	1,232	3,124	3,124
	Convergence rate	100%	100%	69.3%	100%	99.4%	100%
Day 2	Number of stations	3,031	3,031	1,449	1,232	3,031	3,031
	Convergence rate	100%	100%	92.3%	100%	99%	100%
Day 3	Number of stations	3,005	3,005	1,359	1,232	3,005	3,005
	Convergence rate	100%	100%	94.7%	100%	99.3%	100%
Day 4	Number of stations	2,996	2,996	1,342	1,232	2,996	2,996
	Convergence rate	100%	100%	94.8%	100%	99.3%	100%
Day 5	Number of stations	2,994	2,994	1,329	1,232	2,994	2,994
	Convergence rate	100%	100%	95.1%	100%	99.2%	100%
Day 6	Number of stations	2,993	2,993	1,324	1,232	2,993	2,993
	Convergence rate	100%	100%	95%	100%	99.2%	100%

would require data in the extrapolation area, which is not available. In this regard, it is difficult to determine the true value of the 50-year event. As the actual 50-year event is unknown, the only useful method for validating the MRI estimation approaches is to determine the relative level of agreement across methods. Figure 4 displays the relative ratios across the different MRI estimation approaches. These ratios are computed by dividing an x-day design snow load estimate by the median x-day design snow load of the six approaches. This figure combines 50-year MRI estimates from the D-1 to D-6 scenarios, and only shows results for qualifying stations with estimates across the different estimation approaches. The block maxima approach uses zero-inflated probability distributions to account for the fact that some locations have annual maximum snow loads of zero in certain years. With that in mind, we observed that accounting for zero-inflated snow years using precedence set by Buska et al. (2020) had a negligible effect on the MRI estimates. Across the different station types (i.e., FOS, COOP, and SNOTEL), we can see that the traditional block maxima (BM), the mean-adjusted gamma (GMA), and POT Bayesian (BAYES) approach tend to be closer to the reference line of 1, which represents values that equal the median of the six approaches. Figure 4 also shows that the BM approach is conservative since most of the BM estimates fall about the reference line, with relatively little variation in the relative values across stations. The approach's conservative nature may be attractive to structural engineers who find it advantageous to use overestimated design loads in order to avoid structural failures. In contrast, the GMA and BAYES approach exhibits a greater relative variation than the BM approach.

The relative change in the 1-day design snow load estimates for the BAYES and GMA with the BM approach as the reference point is compared in Figures 5. The relative change is computed as the difference between the design load estimate between the compared approach (BAYES and GMA) and the reference approach, divided by the design load of the reference

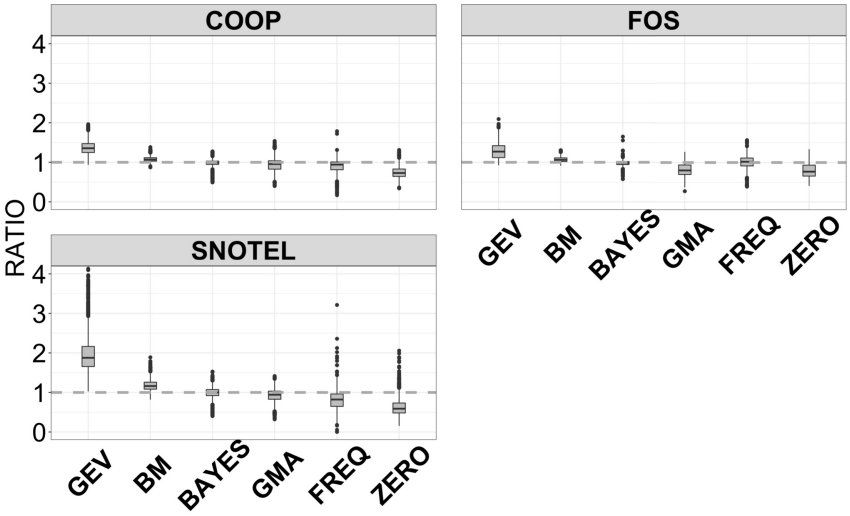
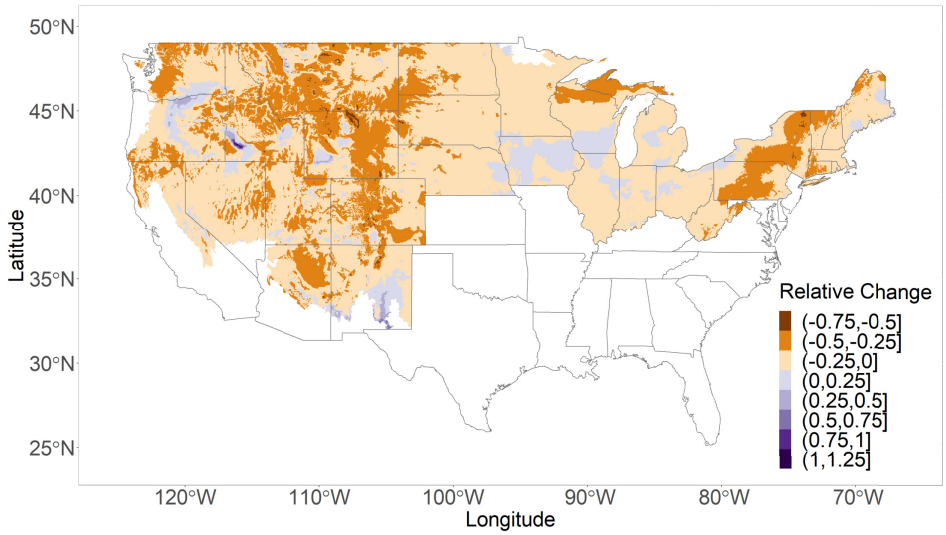


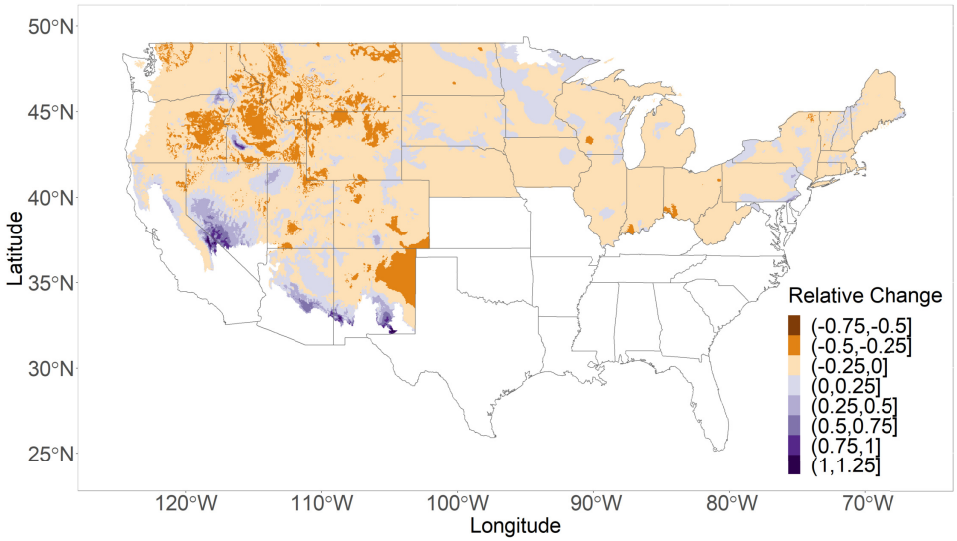
Figure 4: Comparison of relative ratios across the MRI estimation approaches. A relative ratio of 1 (dashed grey line) means that the estimation approach is close to the median of the MRI values for an x-day design snow load estimate. See Table 6 for MRI estimation approaches names and descriptions.

approach. In the spatially continuous map, the layers are created using the RGAM (regional generalized additive model) framework. The framework is accomplished by the remap R package Wagstaff (2021). The response variable is 1-day 50-year event value, while the predictor variables include elevation, mean temperature of the coldest month (30-year average), winter precipitation (30-year average) latitude and longitude, as was used to map the current design snow loads in ASCE 7 (ASCE, 2022) as explained in Bean et al. (2021). The framework creates separate generalized additive models for different geographical regions represented with the 105 level III ecoregions. An ecoregion is an area in the United States partitioned according to similar ecological characteristics. The different prediction surfaces are then stitched together to create a spatially continuous map. However, predictions are made for ecoregions and US states with weather stations greater than 4. Figure 5a shows the BM method tends to produce higher MRI estimates than the GMA method in the Midwest of the US by about 25%. On the west coast and parts of the east coast, the BM method produces design loads that are 50% higher than the GMA method. On the other hand, Figure 5b shows that the BM method predominately produces MRI estimates that are about 25% higher than the BAYES approach.

The POT method has the limitation of sensitivity when it comes to threshold selection. As demonstrated in Figure 7, the sensitive nature of the threshold leads to a higher variance. Based on an x-day Daily Change method, it is expected that the estimated 50-year event will increase or remain the same as the number of days increases. The POT, however, tends to have unexpected behavior regarding some station estimates. Table 8 shows the x-day 50-year event for the Algona weather station in Iowa. The table illustrates that the event estimates for the POT Bayesian tend to increase and decrease as the extracting day increases from day 1 to day 6. This unexpected behavior of non-increasing event value may be due to the selection of different threshold values. As the threshold varies, different exceedances are fitted to the GP distribution. Despite addressing the variance-bias issue with the POT method, the Northrop et al. (2016) approach still leads to substantial variation. This issue applies to Durrieu et al.



(a) Difference plot between the traditional block maxima and GMA approach.



(b) Difference plot between the traditional block maxima and BAYES approach.

Figure 5: Difference plot for the GMA and BAYES approach for 1-day design snow load with the traditional block maxima approach as the reference point. Negative relative change values imply that the BM approach overestimates compared to BAYES or GMA approach.

(2015)'s method, where some weather stations have non-increasing 50-year event values as x-day change increases.

We next examine how design loads using the Daily Change method compare with design loads obtained using season-long accumulations of snow. Figure 6 shows a ratio plot between annual season-long versus an extreme case of the x-day Daily Change method (1-day) on the continental US map.The ratio is computed as the design load value of 1-day over design load of the annual season-long method. The design load is estimated using the block maxima approach for both methods. The mapped values shown in this figure use the same mapping procedure described in conjunction with Figure 5. In Figure 6, we see that the lower ratio values (from

Table 8: 50-year events for Algona weather station in Iowa across the BM, BAYES, and GMA estimation approach. Event values are measured in Kilopascals (Kpa). See Table 6 for MRI estimation approaches names and descriptions.

Day Change	BM	BAYES	GMA
1	0.93	0.89	0.79
2	1.03	1.06	0.94
3	1.05	0.99	1.01
4	1.05	0.98	1.03
5	1.05	0.88	1.04
6	1.05	0.88	1.06

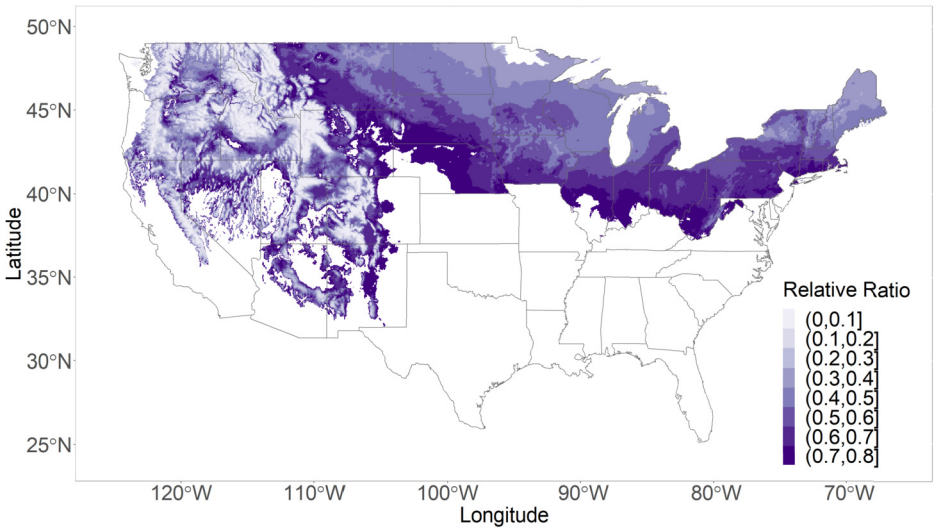


Figure 6: MRI ratio plot of Day-1 block maxima approach to peak, season-long accumulation block maxima approach.

0 to 0.2) tend to congregate in the western part of the country. These low ratio areas tend to be located in the mountainous locations. Figure 6 also shows that the northern part of Minnesota, North Dakota, Michigan, and Maine have a ratio of 0.2 to 0.4. The ratio values tend to increase from 0.2 to about 0.8 as we move from north to south for the Midwest, Mid-Atlantic, and Northeast. This pattern shows that stations farther south across the conterminous U.S. have similar short-term maximum snow load MRIs to their respective traditional design snow load standards. That means that ground-mounted solar panels in the south of the Midwest, Mid-Atlantic, and Northeast may have similar snow load requirements as traditional roofing systems. Thus, using MRI values associated with season-long accumulations of snow for ground-mounted solar panels (GMSP) may lead to overly conservative design requirements in the western U.S. This would result in unnecessarily expensive solar panel mount designs.

This paper considers 1-day MRI to be the most extreme version of short-term snow accumulation. When used in an engineering setting, 1-day design loads may be non-conservative, since it is not guaranteed that there will be sufficient sunlight to shed all the snow within 24 hours in all parts of the United States. The potentially inappropriately low design load obtained using

Table 9: Average MRI change from Day 1 to Day 6 for the BM, GMA, and BAYES approach. R_i - MRI rate of change from Day i to Day i+1, where $i = 1, 2, 3, 4, 5$. See Table 6 for MRI estimation approaches names and descriptions.

		R_1	R_2	R_3	R_4	R_5
BM	FOS	16.9%	7.2%	2.6%	1.4%	0.8%
	COOP	17.8%	6.7%	3.1%	2.1%	0.9%
	SNOTEL	34.6%	17.6%	14.5%	9.7%	6.9%
BAYES	FOS	9.5%	8.8%	4.3%	−0.8%	1.7%
	COOP	14.1%	8.5%	4.6%	4%	1.9%
	SNOTEL	36.9%	21.4%	16.7%	15.2%	9.2%
GMA	FOS	17.9%	4.8%	1.8%	0.8%	0.4%
	COOP	14.5%	5.5%	2.4%	1.1%	0.5%
	SNOTEL	52.7%	25.5%	13.7%	8%	4.6%

1-day loads could result in structural failure. Table 9 illustrates the average increase in MRI for the BM, GMA, and BAYES approaches. In these approaches, the rate of change decreases as the number of days passes from day 1 to day 6. It is generally observed that design snow loads increase consistently for mountainous locations (i.e., SNOTEL) as the x-day window increases, whereas for non-mountainous and southern/mid-latitude locations (i.e., COOP and FOS), the increases tend to level out around day 5 or 6.

5.2 Bootstrap Results

While it is desirable to have a consensus among approaches regarding the estimated 50-year MRI event. It is also desirable to have approaches that are robust to slight changes in the input data. Such changes can be the result of new snow measurements or corrections to measurements in the historical record. In this section, we investigate the performance of the 6 MRI estimation approaches via a bootstrap simulation. In our study, daily snow load measurements x_1, ..., x_N at a given station are bootstrapped N times with replacement. For each qualified station, 100 bootstrap samples are created. In order to obtain bootstrap samples for the traditional block maxima method, maximum daily loads in each snow year are taken from daily bootstrap loads. Next, we use the six estimation approaches to estimate design snow load estimates for the 100 bootstrap samples. The coefficient of variation (COV), which is the standard deviation divided by the mean, is then computed for each estimation method. Figure 7 shows the coefficient of variation plot for each method across station types. A total of 806 weather stations with MRI values across the six estimation approaches were used. The figure shows that the relative size of standard deviation to the mean is lowest for the GMA approach as less emphasis is placed on the tail of a gamma distribution. It can be seen that the two automated threshold selection methods tend to have a high coefficient of variation compared to a fixed threshold at zero. The results highlight the sensitivity of MRI values to automated threshold selection, due to the sensitivity of the GPD parameter estimates to the selected threshold value. This sensitivity could be problematic when trying to use an automatic threshold selection method on thousands of stations across the country as the risk of degenerate distribution fits is higher than some of

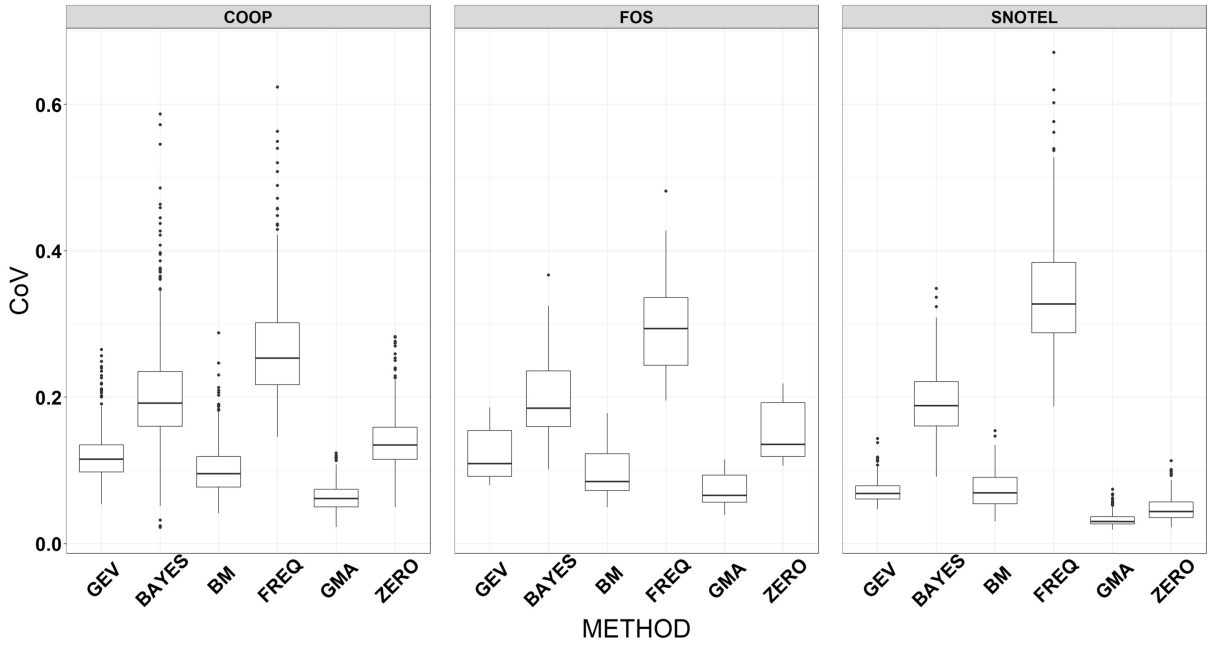


Figure 7: Comparison of the coefficient of variation (CoV) across MRI estimation approaches for a 1-day load. See Table 6 for MRI estimation approach names and descriptions.

the other methods considered.

6 Conclusion

Despite their prevalence, design practices for structures that shed accumulated snow regularly have remained largely unspecified. This paper presented a "storm-level" adaptation of previous structure-related snow studies that is designed to estimate short-term, rather than season-long, accumulations of snow load. Our proposed data extraction process, called the Daily Change method, was based on the concept of sequential blocks of snow accumulations, which considers the sequential consecutive positive differences in the weight of settled snow over an x-day period. The process allowed us to investigate extreme values under various snow-shedding scenarios, such as those observed on ground-mounted solar panels. It has been shown that most non-mountain and non-northern (i.e., FOS, COOP) stations have season-long 50-year MRI snow loads that are similar to the 5-day or 6-day MRI snow loads. This finding reinforces the idea that design standards for structures known to shed snow regularly cannot rely upon traditional design estimates, which assume season-long accumulations of snow on the structure.

This paper provided details on the quality control measures imposed on the data before performing our analysis. This included the development of a random forests model that employed rolling averages of relevant climate variables to estimate missing daily values of SWE or snow weight. These estimates are crucial to supplement the lack of snow weight information at stations that measure only snow depth. We compared six different approaches of extreme value estimation on short-term snow accumulations to illustrate the implications of those approaches as applied to snow load design. It has been observed that the POT Bayesian (BAYES) approach, as well as the mean-adjusted quantile with a gamma distribution (GMA) approach and the traditional

block maxima (BM) approach tend to form a consensus when it comes to estimating 50-year MRI loads (i.e., the estimates are close to one another). However, through a bootstrap study, the POT method exhibited a higher variance (CoV) for slight changes in the input data. The results suggested that the advantage of data efficiency associated with the POT method is invalidated due to the uncertainty in the threshold selection. A future avenue of research may consider ensemble estimates of extreme snow loads using multiple extreme value approaches for estimating these design events. We anticipate that ensemble estimates will prevent the serious consequences of poor x-year MRI estimates of snow loads derived from degenerate distribution fits, which are bound to occur when estimating extreme values at hundreds or thousands of snow measurement locations. Most importantly, the paper has provided a framework for practical comparisons of the efficacy of popularly used approaches for modeling extreme accumulations of snow over short periods of time. This framework will be helpful as we consider the potential evolution of snow accumulation patterns (both in terms of intensity and duration) in a changing climate.

Software

All analysis for this paper was performed in R 4.2.0 (R Core Team, 2022) with the following packages:

- rnoaa (Chamberlain, 2021)
- tidyverse (Wickham et al., 2019)
- snowload2 (Bean et al., 2021)
- extRemes (Gilleland and Katz, 2016)
- fitdistrplus (Delignette-Muller and Dutang, 2015)
- extremefit (Durrieu et al., 2018)
- remap (Wagstaff, 2021)
- threshr (Northrop and Attalides, 2020)

Acknowledgments

The authors would like to thank Dr. Michael O'Rourke, Professor Emeritus at the Rensselaer Polytechnic Institute (RPI), and Patrick Kalush, also affiliated with RPI, whose ongoing work with ground-mounted solar panel design provided motivation for this paper.

References

Abatzoglou JT (2011). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1): 121–131.

ASCE (2022). *Minimum Design Loads and Associated Criteria for Buildings and Other Structures*. American Society of Civil Engineers, ASCE/SEI 7-22 edition.

Balkema AA, De Haan L (1974). Residual life time at great age. *The Annals of Probability*, 2(5): 792–804.

Bean B, Maguire M, Sun Y, Wagstaff J, Al-Rubaye SA, Wheeler J, et al. (2021). The 2020 national snow load study, *Technical Report 276*, Utah State University Department of Mathematics and Statistics, Logan, UT.

Beirlant J, Dierckx G, Guillou A (2005). Estimation of the extreme-value index and generalized quantile plots. *Bernoulli*, 11(6): 949–970.

Bommier E (2014). Peaks-over-threshold modelling of environmental data, *Technical report, U.U.D.M. Project Report 2014:33*, Department of Mathematics, Uppsala University.

Breiman L (2001). Random forests. *Machine Learning*, 45(1): 5–32.

Bücher A, Zhou C (2021). A horse race between the block maxima method and the peak–over–threshold approach. *Statistical Science*, 36(3): 360–378.

Buska J, Greatorex A, Tobiasson W (2020). Site-specific case studies for determining ground snow loads in the United States, *Technical Report ERDC/CRREL SR-20-1*, U.S. Army Engineer Research and Development Center, Cold Regions Research and Engineering Laboratory, Hanover, NH.

Chamberlain S (2021). rnoaa: 'NOAA' Weather Data from R. R package version 1.3.8.

Coles S (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Science & Business Media.

Copland L (2022). 4.04 - Properties of Glacial Ice and Glacier Classification. 52–62. Academic Press, Oxford, second edition.

Curceac S, Atkinson PM, Milne A, Wu L, Harris P (2020). An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales. *Journal of Hydrology*, 585: 124845.

Deidda R (2010). A multiple threshold method for fitting the generalized Pareto distribution to rainfall time series. *Hydrology and Earth System Sciences*, 14(12): 2559–2575.

Delignette-Muller ML, Dutang C (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4): 1–34.

Donegan B (2019). The Most Extreme 24-Hour Rainfall Records in All 50 States. Accessed: 7-26-2022.

Durrieu G, Grama I, Jaunatre K, Pham QK, Tricot JM (2018). extremefit: A package for extreme quantiles. *Journal of Statistical Software*, 87(12): 1–20.

Durrieu G, Grama I, Pham QK, Tricot JM (2015). Nonparametric adaptive estimation of conditional probabilities of rare events and extreme quantiles. *Extremes*, 18(3): 437–478.

Ferreira A, de Haan L (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, 43(1): 276–298.

Fisher RA, Tippett LHC (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: *Mathematical proceedings of the Cambridge philosophical society*, volume 24, 180–190. Cambridge University Press.

Gilleland E, Katz RW (2016). extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software*, 72(8): 1–39.

Gnedenko B (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics*, 44(3): 423.

Jonas T, Marty C, Magnusson J (2009). Estimating the snow water equivalent from snow depth measurements in the Swiss Alps. *Journal of Hydrology*, 378(1–2): 161–167.

Jonathan P, Ewans K (2013). Statistical modelling of extreme ocean environments for marine design: A review. *Ocean Engineering*, 62: 91–109.

McCreight JL, Small EE (2014). Modeling bulk density and snow water equivalent using daily snow depth observations. *The Cryosphere*, 8(2): 521–536.

Menne M, Durre I, Korzeniewski B, Vose R, Gleason B, Houston T (2012). *Global Historical Climatology Network - Daily*. (GHCN-Daily), Version 3.26.

Northrop PJ, Attalides N (2020). threshr: Threshold Selection and Uncertainty for Extreme Value Analysis. R package version 1.0.3.

Northrop PJ, Attalides N, Jonathan P (2016). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(1): 93–120.

Pickands III J (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1): 119–131.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scarrott C, MacDonald A (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1): 33–60.

Sievert C (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Solari S, Egüen M, Polo MJ, Losada MA (2017). Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, 53(4): 2833–2849.

Sturm M, Taras B, Liston GE, Derksen C, Jonas T, Lea J (2010). Estimating snow water equivalent using snow depth data and climate classes. *Journal of Hydrometeorology*, 11(6): 1380–1394.

Thompson P, Cai Y, Reeve D, Stander J (2009). Automated threshold selection methods for extreme wave analysis. *Coastal Engineering*, 56(10): 1013–1021.

Tobiasson W, Greatorex A (1996). Database and methodology for conducting site specific snow load case studies for the United States. In: *Snow engineering: Recent advances: Proceedings of the Third International Conference*, Sendai, Japan, volume 2631, 249256Ye.

Wagstaff J (2021). Regionalized models with spatially continuous predictions at the borders, Ms thesis, Utah State University. Document No. 8065.

Wheeler J, Bean B, Maguire M (2022). Creating a universal depth-to-load conversion technique for the conterminous United States using random forests. *Journal of Cold Regions Engineering*, 36(1): 04021019.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43): 1686.

Yang X, Zhang J, Ren WX (2018). Threshold selection for extreme value estimation of vehicle load effect on bridges. *International Journal of Distributed Sensor Networks*, 14(2): 155014771875769.

Zoglat A, EL Adlouni S, Badaoui F, Amar A, Okou CG (2014). Managing hydrological risks with extreme modeling: Application of peaks over threshold model to the Loukkos Watershed, Morocco. *Journal of Hydrologic Engineering*, 19(9): 05014010.