Block Maxima Methods for Extreme Snowfall Events in Chicago

Patrick Toman*

18 December 2020

Introduction

In the first 9 months of 2020, there have been 16 extreme weather or climate related events with monetary losses exceeding $1 billion dollars in the united states, tying the annual records set by 2011 and 2017 Smith (2020). One of the most disruptive events, particularly for urbanized areas, are extreme snow events. One question at the forefront of researchers and planners minds is whether or not the frequency and intensity of extreme winter storms is increasing. To that end, Lee and Lee (2020) proposed methods based on *generalized extreme value distributions* for modeling trend and return levels for extreme snow events in New York City using 56 years of annual snowfall data. They found that their modeling framework was useful in modeling explaining extreme snow events in the city. Given Lee and Lee's reported success, an obvious line of inquiry would be to assess if this model can be applied to a different geo-spatial location. With this in mind, this report takes the methods from Lee and Lee's paper and applies them to annual snowfall data for Chicago, another city that is prone to crippling blizzards.

Data

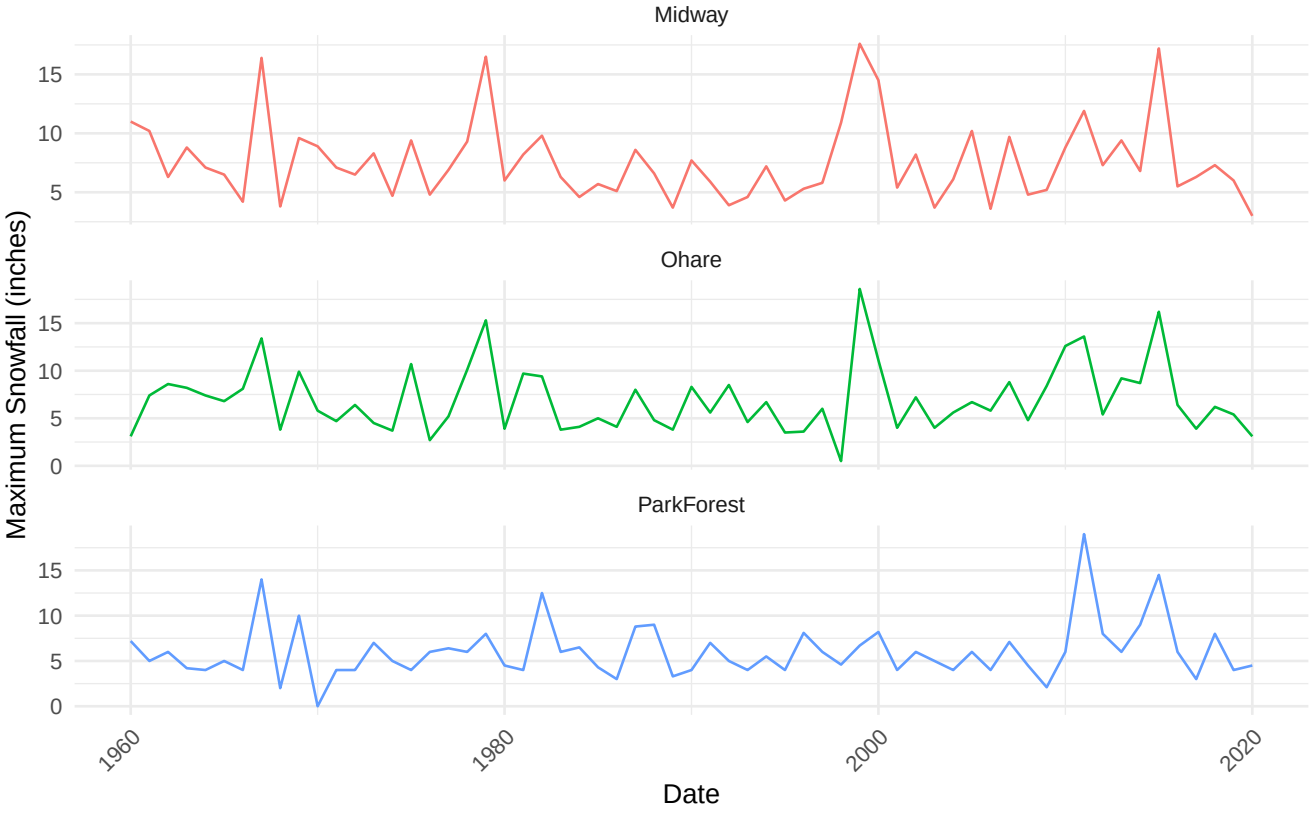
Daily snowfall data is downloaded from the NCEI (2020) using the Climate Online Data (CDO) tool. Three weather stations are selected from the Chicago Area: Midway, O'Hare, and Park Forest. More geographic details can be found in table 1 below. The last 61 years of daily snowfall data from each of the three weather stations are used in the analysis, starting from the dates July 1st,1960 to June 30th, 2020. Daily snowfall is defined as the maximum amount of that has accumulated prior to melting or settling for the day. The snowfall data from NCEI is measured in inches with the amounts being rounded to the nearest tenth of an inch with amounts less than 0.1 being recored as zeros. Furthermore, $\approx 8\%$ of the days in our analysis have non-trace snow amounts. We refer to these days with non-trace snowfall as *snow events*. Since major snowfall events tend transpire over the course of several, consecutive observations with non-zero snowfall are merged to represent one single snow event,thus, a snowfall observation refers to the accumulated snowfall associated with a given storm. Finally, a *snow year* is defined as a one-year period that starts on July 1st to June 30th. For instance, the snow records from July 1st 2012 and June 30th, 2013 would correspond to the 2012 snow year. Note that there are missing observations in the data. For Midway, $< 0.8\%$ of daily observations are missing, O'Hare is missing $\approx 0.9\%$ of observations and $\approx 8\%$ are missing for Park Forest. Most of the missingness appears to be concentrated in non-winter months, therefore, the issue of missing data is negligible. Figure 1 , shows a plot of the annual maximum snowfall at each station for the time period starting July 1st, 1960 and ending June 30th, 2020.

*patrick.toman@uconn.edu; Ph.D. student at Department of Statistics, University of Connecticut.

Table 1: Station Summaries

Station	Full Station Name	Latitude	Longitude	Elevation
Chicago	CHICAGO MIDWAY AIRPORT 3 SW, IL US	41.74	-87.78	189.00
Park Forest	PARK FOREST, IL US	41.49	-87.68	216.40
O'Hare	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	41.96	-87.93	201.80

Figure 1 − Annual Maxes



Methods

Block Maxima Methods for GEV

Let $X_1, \ldots, X_k$ be independent and identically distributed random variables that have the common CDF $F(.)$. Next, let us define $M^{(k)} = max\{X_1, \ldots, X_k\}$ as the maximum order statistic for a *block* of k these random variables. Suppose then that there are a set of constants $\{a_k\}$ and $\{b_k\}$ with $b_k > 0\ \forall k$ such that

$$P\left(\frac{M^{(k)} - a_k}{b_k} \leq x\right) \to G(x) \text{ for } k \to \infty \tag{1}$$

where $G(.)$ is a non-degenerate distribution function. Then according to the Fisher-Tippett-Gnedenko theorem, $G(.)$ follows one of these distribution families: Gumbel,Frechet, or Weibull. These three families can then be further generalized into the *generalized extreme value*(GEV) family of distributions where we have

$$G(x) = exp\left\{-\left[1+\xi(\frac{x-\mu}{\sigma})\right]_+^{\frac{-1}{\xi}}\right\} \tag{2}$$

where $z_+ = max\{z, 0\}$, and we have unknown real valued parameters $\mu \in \mathcal{R}, \sigma \in \mathcal{R}^+$, and $\xi \in \mathcal{R}^+$ referred to as location,scale,and shape parameters,respectively. In applied settings, block maxima methods rely on a sequence of of maximum order statistics from a CDF $F(.)$. If we have a random sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ then we can attain a series of block maxima from this data by dividing $\mathbf{X}$ into k non-overlapping blocks and then finding the maximum within each block to attain $\mathbf{M} = \{M_1, \ldots, M_k\}$, a k-dimensional set of block maxima. If the block sizes are large enough, then the GEV family of that In our case, each snow year is treat as a block and then we extract the maximum snowfall statistic from each snowfall year for each station as our extreme event for that year. In our particular case, we treat each snow year as a block, thus, $M_i,\ i = 1960, \ldots, 2020$ is the maximum snowfall event for a given snow year.

Maximum Likelihood for GEV Models

Assume that we have $\xi \neq 0$. If $X_1, \ldots, X_N \overset{iid}{\sim} GEV(\Theta)$ where $\Theta = (\mu, \sigma, \xi), -\infty < \mu, \xi < \infty, \sigma > 0$ is our vector of unknown parameters. then we have the log-likelihood for for $\mathbf{X} = \{X_1, \ldots, X_N\}$ as

$$ln\left(\Theta|\mathbf{X}\right) = -Nln(\sigma) + \left(\frac{1}{\xi}+1\right)\sum_{i=1}^{N}\left(1\xi\left(\frac{x_i-\mu}{\sigma}\right)\right) - \sum_{i=1}^{N}\left(1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right) \tag{3}$$

assuming that $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$. Further information can be found in Beirlant et al. (2005). Of course, parameter estimates can be obtained by numerical estimation methods such as Newton-Raphson. In this analysis, we use the standard *nlme()* function found in the R programming language for maximum likelihood estimation.

Return Levels

Some of the most important quantities in any extreme values analysis are the return levels. The return level for an associated return period of K years is the expected level that is to be exceeded on average once over the following K years. In the case of annual maximums we have the return level X_k for a period of K years defined as

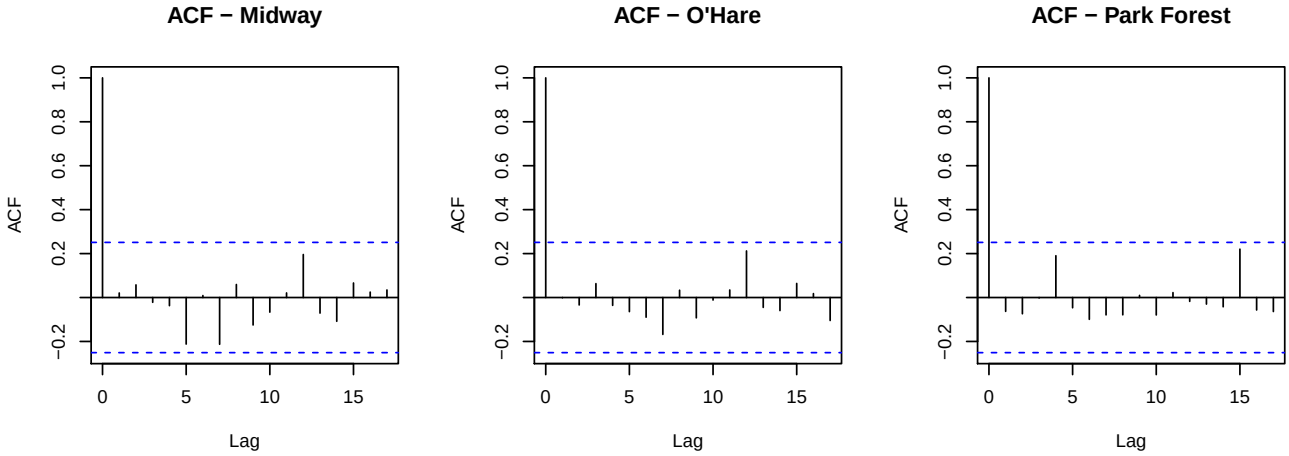
$$X_k = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-ln(1-K^{-1})\}^{-\xi}\right] & \text{if } \xi \neq 0 \\ \mu - \sigma\left[-ln(1-K^{-1})\right] & \text{if } \xi = 0 \end{cases} \tag{4}$$

Invoking the invariance property of MLEs, we can solve for the MLE of the return level $\hat{X}_k$ by simply substituting the MLEs for (μ, σ, ξ) attained through Newton-Raphson.

Standard Error Correction for Spatial and Temporal Dependence

Snowfall data collected from stations close in proximity to one another will typically exhibit spatial correlation. To account for this spatial correlation in our standard errors, we can employ a method devised by Smith Smith (1990) where MLE estimates are attained under the standard IID assumption and then standard errors are corrected in a post-hoc fashion to account for spatial dependence. More information about this method can be found in the appendix. Note that for block maxima methods, temporal dependence is typically not an issue since annual maxima tend to be separated by a large gap of

time. To confirm this,we calculate the ACF of the annual maxima series for each station and find that there is no statistically significant auto-correlation for the $\alpha = 0.05$.



Bootstrap CI for Return Levels

For GEV models, Rust (2011) note that asymptotic standard errors attained via the delta method are often inadequate in estimating the sampling variability of MLE estimators for return levels. Lee and Lee (2020) propose bootstrapping the confidence intervals for return levels using the standard percentile method where we use the $\alpha/2$ upper and lower quantiles of the bootstrapped return level estimates. Because the distribution of return levels tends to exhibit right-skew, therefore, standard bootstrap methods will be biased. To remedy this bias in bootstrap return level estimates, we can employ the *bias corrected and accelerated* (BCa) bootstrap methods introduced by Efron (1987). Suppose that we wish to generate B bootsrap estimates of return level X_k. The BCa method works by first calculating the bias correction estimate z_{BC}

$$z_b c = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I\left(\hat{x}_K^{(b)} < \hat{x}_K\right)\right) \tag{5}$$

where $\Phi(.)$ is the standard normal CDF, $I(.)$ is the indicator function, and $\hat{x}_k$ denotes the MLE of the return level for period K using the original data. Next, we calculate the acceleration constant c_A which is defined as

$$c_A = \frac{\sum_{t=1}^{n}\left(\tilde{x}_K^{-t} - \tilde{x}_K\right)^3}{6\left[\sum_{t=1}^{n}\left(\tilde{x}_K^{(-t)} - \tilde{x}_K\right)^2\right]^{3/2}} \tag{6}$$

where $\tilde{x}_K^{(-t)}$ denotes the delete-1 jacknife estimate of x_K where we have deleted the t^{th} observation from the dataset and $\tilde{x}_K^{(-t)} = \sum_{t=1}^{n} \tilde{x}_K^{(-t)}$. Further information can be found in Givens and Hoeting (2013). Thus, the $(1-\alpha) * 100\%$ BCa is interval has the following quantiles as its upper (lower) endpoints given by

$$\Phi\left(z_{BC} + \frac{z_{BC} \pm z_{\alpha/2}}{1 - c_A(z_{BC} \pm z_{\alpha_2})}\right) \tag{7}$$

Implementation and Results

Using the notation given by Lee and Lee (2020), let N_s denote the number of snowstorms that occured at station s during the study period from July 1st, 1960 to June 30th, 2020. If we let $\mathbf{X} = \{X_{s,1}, \ldots, X_{s,N_s}\}$ represent the the accumulated snowfall observed at station s where $X_{s,j},\ j \in 1, \ldots, N_S$ denotes the accumulated snowfall for snowstorm j during the study period. Next, if we denote $M_{s,t}$ denote the maximum for each snow year $t \in \{1960, \ldots, 2020\}$, then we have $M_{s,t} \sim GEV(\mu_s, \sigma_s, \xi_s)$. Using the block maxima methods described in the Methods section, we fit several different models and compare them using Akaike's Corrected Information Criterion (AICc) where $AICc = 2\left(p - \ell(\hat{\Theta}) + \frac{p(p+1)}{n-p+1}\right)$ where p is the number of parameters, $\ell(\hat{\Theta})$ denotes the value of the log-likelihood evaluated at the MLE, and n is the number of observations. Models with smaller AICc are preferred. In addition, we calculate $100 \cdot (1-\alpha)\%, \alpha = 0.05$ CI BCa intervals for return levels $X_k, K = \{25, 50, 75, 100\}$ using $B = 10000$ replications for all six reduced models mentioned in table **??**.

Models

Initially, we build a full model in which we assume that each station has its own location, shape, and scale parameters. In other words, we have $\Theta_s = (\mu_s, \sigma_s, \xi_s), s = 1, 2, 3$. The AICc of this model is -884.1794. Next, we build a model where μ_s is fit individually for each station but σ and ξ are assumed to be the same for all three locations. For the reduced model, we find that the AICc is -895.6098. Therefore, we prefer the reduced model where μ_s's are assumed to be different for the stations but σ and ξ are assumed common.

Having established that σ and ξ should be common to all three stations, three different stationary models, one with a distinct μ_s for all three locations, the second where we merge the two closest μ_s's from the first model, and then a final model where μ_s is assumed to be the same across all three equations. In all three cases, maximum likelihood estimates are obtained using Newton-Raphson via the *nlme()* function in R. Naive standard errors are obtained for these models by inverting the hessian returned from *nlme()*. To account for spatial correlation in the standard errors, we implement Smith's method. Indeed, we find that the corrected standard errors are larger than the naive standard errors, indicating that there is some measure of spatial correlation in the annual maxima series as anticipated. Observe that the estimates for ξ are quite small relative to their standard error with the largest $\frac{\hat{\xi}}{se(\hat{\xi})} \approx 1$. Appealing to asymptotic normality results, we can conclude that ξ is not statistically different from 0 at the $\alpha = 0.05$ level. More details can be found in table 2.

Next, because of the possibility of a a linear trend in the data, a set of non-stationary models are fit following the same structure as the stationary models except we now model the location parameters μ_s as a function of time. Following the example of Lee and Lee (2020), we have

$$\mu_{s,t} = \mu_s + \beta\left(\frac{t - 1960}{10}\right),\ t = 1960, \ldots, 2020 \tag{8}$$

where β is a trend parameter that captures the expected in maximum snowfall over a decade. It is assumeth that trend is the same for all three stations in our analysis. The non-stationary parameter estimates are quite similar to the stationary estimates. In addition, by invoking asymptotic properties of the MLE, we can conclude that trend parameter β is not statistically different from zero at the $\alpha = 0.05$ level, thus, we can reasonably conclude that there is not a long-term trend in the maximum snowfall for Chicago in this time frame. Once again, more details are given in the table 2 below.

Table 2: GEV Estimates W/ Standard Errors in Parentheses (left: Uncorrected, right: Corrected)

	Stationary Models			Non-Stationary Models		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
μ_M	6.150 (0.308,0.330)	6.152 (0.309,0.639)	5.367 (0.201,0.246)	6.116 (0.462,0.498)	6.079 (0.463,0.734)	5.214 (0.389,0.462)
μ_O	5.242 (0.321,0.370)	4.978 (0.234,0.272)	5.367 (0.201,0.246)	5.210 (0.457,0.484)	4.914 (0.389,0.598)	5.214 (0.389,0.462)
μ_{PF}	4.743 (0.311,0.337)	4.978 (0.234,0.272)	5.367 (0.201,0.246)	4.716 (0.421,0.538)	4.914 (0.389,0.598)	5.214 (0.389,0.462)
β	N/A	N/A	N/A	0.01 (0.102,0.128)	0.022 (0.101,0.154)	0.048 (0.105,0.133)
σ	2.392 (0.141, 0.224)	2.398 (0.141,0.293)	2.485 (0.142,0.241)	2.392 (0.141,0.224)	2.397 (0.141,0.295)	2.481 (0.144,0.238)
ξ	0.045 (0.046, 0.056)	0.046 (0.046,0.087)	0.028 (0.044,0.056)	0.045 (0.461,0.056)	0.048 (0.046,0.088)	0.030 (0.044,0.056)
ℓ	452.974	453.663	458.117	452.980	453.640	458.013
$AICc$	-895.609	-899.101	-910.101	-893.462	-896.942	-907.800

Bootstrap Return Levels

We present results on the BCa return levels for both the stationary and non-stationary versions of model 1. Model 1 is chosen since there are not substantial differences in the AICc values for the models, however, there does seem to be some differences in the location the location parameters that are worth exploring. First, we observe that the 95% CI is fairly similar for both models though it is a bit wider for the trend model. This helps to corroborate our findings that the trend parameter β is not statistically different from 0, thus, indicating that there is no long term trend in the annual maximum snowfall. Furthermore, the CI plots indicate that we can expect a snowstorm that dumps ≈ 15 inches of snow or more occurs once every 25 years or so in the chicago area. When we compare this to the actuall data, we do see that such snowstorms do seem to occurs once every 25 years or so. Finally, we observe that the return median return levels for all periods are higher for Midway than for the other two sites.

Conclusions and Further Work

In this report, we investigated whether or not the block maxima methods presented by Lee and Lee (2020) could be applied to a different dataset of annual maximum snowfall series from Chicago. Overall, we corroborate many of the same findings for the Chicago area. Foremost, we find that there is indeed a fair amount of spatial correlation in the data, thus demonstrating th need for application of Smith's method. Furthermore, we find that for Chicago there is not a statistically significant trend in the annual maximum snowfall event across the study period. Finally, we found that the bias corrected bootstrap is effective in giving us more accuarate assessments of the uncertainty in K-year return levels.

The analysis in this report also helps to reveal further possible avenues of inquiry. First, Lee and Lee also presented in their work a threshold exceedance model where snowfall events exceeding a certain threshold are modeled using a generalized pareto distribution. Such a model helps to add information by incorporating large snowfall events other than just the annual maximum. Applying this method to the Chicago data may reveal new findings about the nature of blizzards in the Chicago area. In addition, it would be interesting to apply this method to other extreme weather data such as wind speeds, rainfall, and temperatures. Finally, our model only incorporated a trend covariate for a subset of the models, it would be prudent to investigate if any other covariates have a significant relationship with extreme snowfall events.

Appendix

Smith's Method

Suppose that we have multiple weather stations in the same region, each with t years of recorded snowfall. The log-likelihood for the vector of unknown parameters $\Theta = (\theta_1, \ldots, \theta_p)^T$ using all stations' data as for all years t can be expressed as

$$\ell_t(\Theta) = \sum_{i=1}^{t} h_t\left(\Theta\right)$$

where $h_i(\Theta)$ is the contribution of to the log-likelihood of all s stations for the t^{th} year, independent of the other t years. Let $\hat{\Theta} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_p\right)^T$ denote the maximum likelihood estimate of Θ and let Θ_0 denote the true value of Θ. Taylor expansion of the log-likelihood yields

$$\hat{\Theta} - \Theta_0 \approx \left\{-\nabla^2 \ell_t(\Theta_0)\right\}^{-1} \nabla \ell_t(\Theta_0)$$

where *nabla* denotes the gradient and ∇^2 denotes the hessian. We can approximate the entries of the hessian matrix using expected values, thus we have

$$cov(\hat{\Theta}) \approx H^{-1} V H^{-1}$$

where $H = -E\left[\nabla^2 \ell_t(\Theta_0)\right]$ and $V = cov\left(\nabla \ell_t(\Theta_0)\right)$. Assuming that there was no spatial correlation in the data, we would simply have

$$cov(\hat{\Theta}) \approx H^{-1}$$

where H is approximated by the *observed fisher information* $-\nabla^2 \ell_t(\hat{\Theta})$. Furthemore, suppose that the time series data that contribute to $\ell_t(\Theta)$ are dependent but each $h_i(\Theta)$ are independent. In addition, assume that $h_i(\Theta)$ share a common distribution. Unde these two asumptions, we can express the gradient of $\ell_t(\Theta)$ as a sum of independent terms as follows

$$\nabla \ell_t(\Theta) = \sum_{i=1}^{t} \nabla h_i(\Theta)$$

The covariance matrix of $\nabla h_i(\Theta_0)$ can be approximated by the empirical covariance matrix of $\nabla h_i(\hat{\Theta})$, $\forall\, i \in \{1, \ldots, t\}$ and setting.

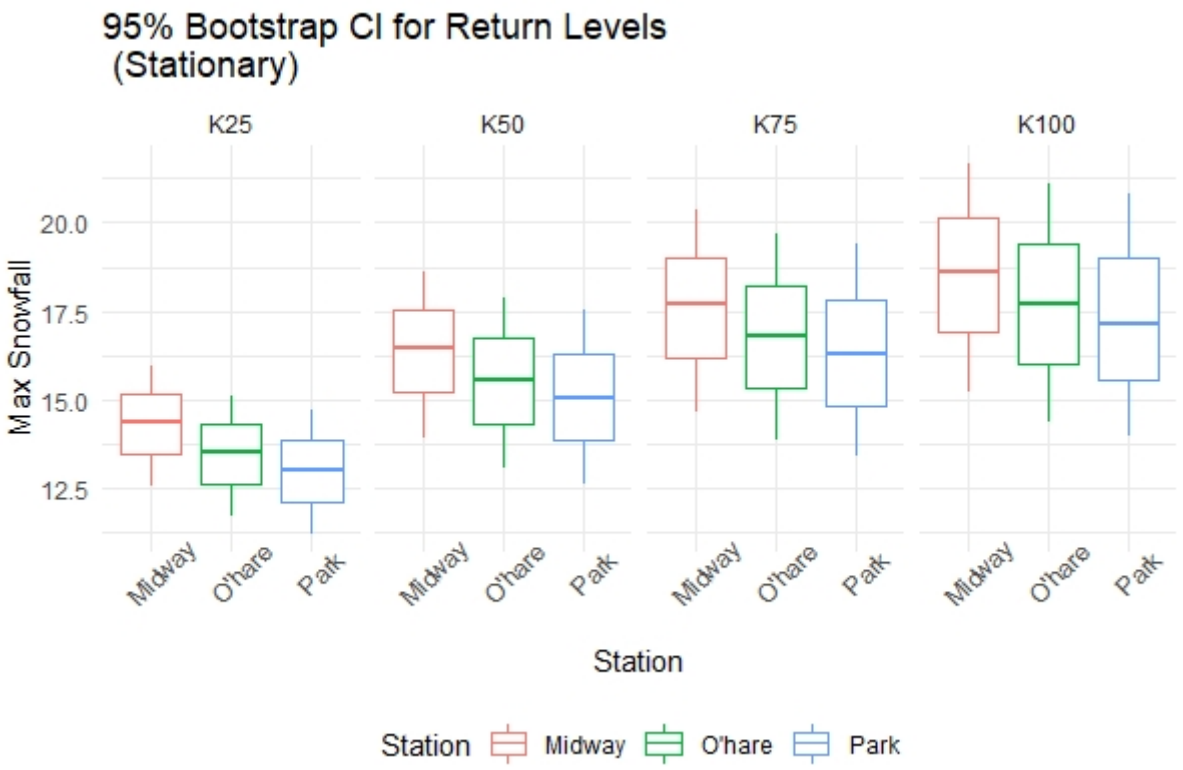
$$V = n \nabla h_1(\Theta)$$

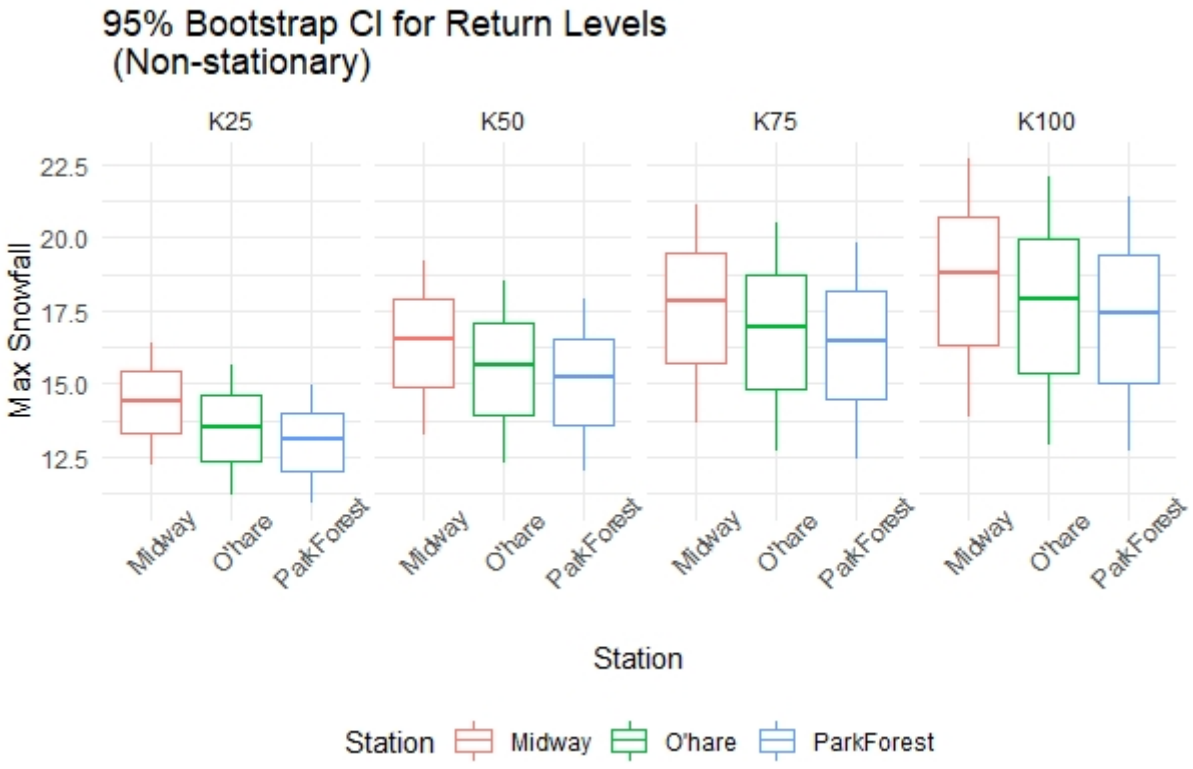
If we combine this expression with the observed information matrix as an estimator of H, we can recover the expression

$$cov(\Theta) = H^{-1} V H^{-1}$$

in a computationally simple manner. More details are given by Smith (1990).

BCa Return Level CI





References

Beirlant, Jan, Jozef Teugels, Johan Segers, and Yuri Goegbeur. 2005. *Statistics of Extremes: Theory and Applications*. Wiley.

Efron, Bradley. 1987. "Better Bootstrap Confidence Intervals." *Journal of the American Statistical Association* 82 (397): 171–85. http://www.jstor.org/stable/2289144.

Givens, Geof H., and Jennifer A. Hoeting. 2013. *Computational Statistics*. Wiley.

Lee, Mintaek, and Jaechoul Lee. 2020. "Trend and Return Level of Extreme Snow Events in New York City." *The American Statistician* 74 (3): 282–93. https://doi.org/10.1080/00031305.2019.1592780.

NCEI. 2020. "Climate Data Online." *Climate Data Online (CDO) - the National Climatic Data Center's (NCDC) Climate Data Online (CDO)*. https://www.ncdc.noaa.gov/cdo-web/.

Rust, Malaakand Schellnhuber, Henning W.and Kallache. 2011. "Confidence Intervals for Flood Return Level Estimates Assuming Long-Range Dependence." In *In Extremis: Disruptive Events and Trends in Climate and Hydrology*, edited by Jürgen Kropp and Hans-Joachim Schellnhuber, 60–88. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14863-7_3.

Smith, Adam. 2020. "NOAA National Centers for Environmental Information (Ncei) U.s. Billion-Dollar Weather and Climate Disasters." https://www.ncdc.noaa.gov/billions/.

Smith, R. L. 1990. *Regional Estimation for Spatially Dependent Data*.