The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Trend and Return Level of Extreme Snow Events in New York City

Mintaek Lee & Jaechoul Lee

To cite this article: Mintaek Lee & Jaechoul Lee (2020) Trend and Return Level of Extreme Snow Events in New York City, The American Statistician, 74:3, 282-293, DOI: 10.1080/00031305.2019.1592780

To link to this article: https://doi.org/10.1080/00031305.2019.1592780

THE AMERICAN STATISTICIAN
2020, VOL. 74, NO. 3, 282–293: Statistical Practice
https://doi.org/10.1080/00031305.2019.1592780

Trend and Return Level of Extreme Snow Events in New York City

Mintaek Lee and Jaechoul Lee

Department of Mathematics, Boise State University, Boise, ID

ABSTRACT

A major winter storm brought up to 42 inches of snow in parts of the Mid-Atlantic and Northeast states for January 22–24, 2016. The blizzard of January 2016 impacted about 102.8 million people, claiming at least 55 lives and $500 million to $3 billion in economic losses. This article studies two important aspects of extreme snowfall events: 1. trends in annual maxima and threshold exceedances and 2. return levels for extreme snowfall. Applying extreme value methods to the extreme snow data in the New York City area, we quantify linear trends in extreme snowfall and assess how severe the 2016 blizzard is in terms of return levels. To find a more realistic standard error for the extreme value methods, we extend Smith's method to adapt to both spatial and temporal correlations in the snow data. Our results show increasing, but insignificant trends in the annual maximum snowfall series. However, we find that the 87.5th percentile snowfall has significantly increased by 0.564 inches per decade, suggesting that, while the maximum snowfall is not significantly increasing, there have been increases in the snowfall among the larger storms. We also find that the 2016 blizzard is indeed an extreme snow event equivalent to about a 40-year return level in the New York City area. The extreme value methods used in this study are thoroughly illustrated for general readers. Data and modularized programming codes are to be available online to aid practitioners in using extreme value methods in applications.

ARTICLE HISTORY
Received August 2017
Revised December 2018

KEYWORDS
Bootstrap confidence interval; Generalized extreme value distribution; Generalized Pareto distribution; Spatial and temporal correlation

1. Introduction

Extreme weather and climate events greatly impact human beings, societies, and ecosystems. In the United States, there have been 203 weather and climate related disasters from 1980 to 2016, each with at least $1 billion in damages (NCEI 2017). Among these extreme weather and climate events, extreme snow events can disastrously affect urban life in particular. Recently, the blizzard of January 2016 brought an all-time record snowfall of 26.8 inches in Manhattan, resulting in significant damages/losses to many urban areas and life in the northeastern United States (National Weather Service 2016). The 2016 blizzard raises an important question: Are we experiencing disastrous snowstorms more often than before? In contrast, many authors have reported decreasing trends in snowfall in various regions. Kunkel et al. (2009) found that snowfall is strongly decreasing in the Mid-Atlantic coast. Burakowski et al. (2008) obtained a decreasing trend of 1.81 inches decade^{-1} in average winter snowfall in the northeastern United States. Huntington et al. (2004) verified a statistically significant decreasing trend in the ratio of snow to total precipitation (S/P) in the northeastern United States, commenting the decreasing trend in S/P is mostly related to decreasing snowfall. According to these findings, the 2016 blizzard appears to be arguably contradictory to a common perception of global climate change: increasing temperature reduces snowfall.

We consider less-studied—but very critical to our life—questions, such as: are extreme snow events changing or not? If extreme snow events change, are they decreasing or increasing? How severe would the snow events we experience be in the next 25 or 50 years? One should not necessarily conjecture that extreme snow events would show the same pattern as average snow events. In fact, extreme and mean statistics are statistically independent in large samples under some mild conditions (McCormick and Qi 2000). This implies that the analysis of extreme statistics can exhibit different results from that of mean statistics. Relevant to our example, O'Gorman (2014) showed that for most land regions in the northern hemisphere, changes in extreme snowfall were very small even though mean snowfall decreased significantly based on climate simulations under high carbon dioxide emission scenarios.

Extreme data analysis requires the use of extreme value methods. Since extreme data usually do not follow the Gaussian distribution, the use of Gaussian-based methods does not produce accurate results for extreme data. With respect to extreme weather events, Fawcett and Walshaw (2007) used the generalized Pareto (GP) distribution to model extreme sea surge heights in Newlyn, UK and compared methods in temporally dependent settings. Northrop and Jonathan (2011) used a quantile regression technique to model nonstationary thresholds of their extreme value model for hurricane-induced wave heights in the Gulf of Mexico. Rust et al. (2011) used the generalized extreme value (GEV) distribution and bootstrap to estimate confidence intervals of return levels for floods in the southern Germany. Fawcett and Walshaw (2012) used extremal index estimation and bootstrap to obtain confidence

CONTACT Jaechoul Lee *jaechoullee@boisestate.edu* Department of Mathematics, Boise State University, Boise, ID 83725.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/TAS*.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TAS*.

intervals of return levels for sea surge heights in Newlyn, UK and wind speeds in Bradfield, UK. Lee, Li, and Lund (2014) used GEV distribution with nonstationary location parameters and a changepoint technique to quantify linear trends for maximum and minimum temperatures in the contiguous United States. Panagoulia, Economou, and Caroni (2014) used GEV and model selection criteria to analyze extreme precipitation in a mountainous Mesochora catchment in Greece.

Some researchers have used extreme value methods to analyze extreme snowfall events. Makkonen et al. (2007) applied climate simulations and GEV to compute 50-year return levels for snowfall in the Nordic area. Blanchet et al. (2009) used a Poisson point process representation to analyze extreme snowfall data collected in the Swiss Alpine region. López-Moreno et al. (2011) employed climate simulations and GP distribution to analyze extreme snowfall in the Pyrenean mountain range located on the border of Spain and France.

We aim to study two important aspects of extreme snow events in an urban area. First, we quantify trends in annual maxima and threshold exceedances of daily snowfall data observed in the New York City area. Second, we specify and estimate return levels for these extreme snow events. This paper uses recent statistical methods based on extreme value theory. To obtain more realistic standard errors, we also extend Smith's method (Smith 1990) to adapt to both spatial and temporal correlations in snowfall data. We thoroughly illustrate these extreme value methods for practitioners in diverse areas.

The rest of this paper proceeds as follows: Section 2 describes the snowfall data used in our analysis. Section 3 explains the extreme value theory and methods applied to the data. Section 4 summarizes our findings on extreme snow events. In Section 5, we conclude with further comments.

2. The Data

The snowfall data in this study were downloaded from the website of the Northeast Regional Climate Center (NRCC) in Ithaca, New York. We selected four weather stations from the New York City area: Central Park, Newark, La Guardia, and JFK. Table 1 shows the geographical descriptions of these four weather stations. We used the most recent 56 years of the NRCC's daily snowfall data of the four stations, starting from July 1, 1959 and ending on June 30, 2015.

Snowfall for a day is the maximum amount of new snow and ice that have accumulated prior to melting or settling for the day. NRCC's snowfall measurements are recorded in inches and rounded to the nearest tenth of an inch. The snowfall measurements of less than 0.1 inches are typically recorded as "trace." We treat these nearly zero snowfall amounts as zero inches, which results in 3.9% of daily snowfall data as meaningful snowfall records (≥ 0.1 inches). Although the NRCC data contain daily observations, many extreme snowfall events tend to last more than one day, producing multiple daily snowfall observations from a single snowstorm. For example, a snowstorm in 2010 brought Central Park 9.4 inches of snow on February 25 and 11.5 inches of snow on February 26, totaling 20.9 inches of snow for the two consecutive days. To properly assess the snowfall amount from each snowstorm, we merged daily nonzero snowfall observations within each continuous snow event. If consecutive snowfall observations of greater than zero inches are not separated by a snowfall record of zero inches, we consider them to be from the same snowstorm. In our analysis, snowfall observation refers to accumulated snowfall associated with each snowstorm.

Table 1. Geographical information on the selected weather stations.

Station	Full station name	Latitude	Longitude	Elevation
Central Park	NEW YORK CNTRL PK TWR	40.7789°	-73.9692°	132 ft
Newark	NEWARK LIBERTY INTL AP	40.6825°	-74.1694°	29 ft
La Guardia	LA GUARDIA AP	40.7794°	-73.8803°	39 ft
JFK	JFK INTL AP	40.6386°	-73.7622°	11 ft

The selected NRCC snowfall data contain missing values, but missing rates are very low. For the study period of 1959–2016, La Guardia has no missing days, Central Park and Newark have only two missing days, and JFK has seven missing days. We exclude all missing data from our analysis, since the missing rates are minimal.

We treat a "snow year" as a one-year period from July 1 to June 30. For instance, the snow records from July 1, 1999 to June 30, 2000 are treated as the observations in the snow year of 1999. The observations up to June 30, 2015 are first used in our analysis to determine if our extreme value methods could adequately predict the 2016 blizzard. Results from partial data up to June 30, 2015 are later compared to the results from the full data up to June 30, 2016, which includes those extreme observations from the January 2016 blizzard.

3. Methods

3.1. Block Maxima Methods

Suppose that $X_1, \ldots, X_k$ are independent and identically distributed (IID) random variables taken from a population with a common distribution function $F(\cdot)$. Define $M^{(k)} = \max\{X_1, \ldots, X_k\}$ as the maximum statistic for a set (block) of these k random variables. If there are sequences of constants $\{a_k\}$ and $\{b_k\}$ with $b_k > 0$ that scale $M^{(k)}$ such that

$$P\left(\frac{M^{(k)} - a_k}{b_k} \leq x\right) \longrightarrow G(x) \quad \text{as } k \to \infty,$$

where $G(\cdot)$ is a nondegenerate distribution function, then $G(\cdot)$ belongs to one of the following extreme value distribution families:

$$\text{(Gumbel)} \quad G(x) = \exp\left\{-\exp\left[-\left(\frac{x-a}{b}\right)\right]\right\}, \quad -\infty < x < \infty,$$

$$\text{(Fréchet)} \quad G(x) = \begin{cases} 0 & \text{if } x \leq a; \\ \exp\left\{-\left(\frac{x-a}{b}\right)^{-\alpha}\right\} & \text{if } x > a, \end{cases}$$

$$\text{(Weibull)} \quad G(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-a}{b}\right)^{\alpha}\right]\right\} & \text{if } x < a; \\ 0 & \text{if } x \geq a \end{cases}$$

for some constants a, $b > 0$, and $\alpha > 0$ (Fisher–Tippett–Gnedenko theorem). These families can be further unified into

the following generalized extreme value (GEV) distribution function:

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

where $z_+ = \max\{z, 0\}$, $-\infty < \mu < \infty$, $\sigma > 0$, and $-\infty < \xi < \infty$. The unknown constants μ, σ, and ξ are called location, scale, and shape parameters, respectively.

In practice, block maxima methods use a sequence of these maximum statistics. Suppose $X_1, \ldots, X_N$ are a random sample from a distribution function $F(\cdot)$. To obtain block maximum series for this sample, we take a maximum statistic within each of n preset blocks and denote the tth block maximum statistic as M_t for $t = 1, \ldots, n$. The previous extreme value theorem implies that if block size is large enough, the GEV distribution can be an approximate probability distribution for the maximum statistics $\{M_1, \ldots, M_n\}$ regardless of the distribution function $F(\cdot)$ from which the original sample $\{X_1, \ldots, X_N\}$ is taken. In practice, the maximum statistics M_t is taken as the maximum of data values recorded for a meaningful period, most often years. The GEV parameters are numerically estimated by the maximum likelihood method.

Return levels are crucial quantities in extreme value studies. The return level associated with the return period of K years is the expected level that is to be exceeded on average once over following K years. For annual maximum data, the K-year return level x_K is

$$x_K = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\ln(1-K^{-1})\}^{-\xi}\right], & \text{if } \xi \neq 0; \\ \mu - \sigma\ln\{-\ln(1-K^{-1})\}, & \text{if } \xi = 0. \end{cases} \quad (1)$$

Substituting the GEV parameters μ, σ, and ξ in (1) with their maximum likelihood estimates finds the maximum likelihood estimate for the return level x_K.

3.2. Threshold Exceedances Methods

Suppose $X_1, \ldots, X_N$ are a random sample from a distribution function $F(\cdot)$. Let X be any of these random variables, and assume $F(\cdot)$ satisfies the Fisher–Tippett–Gnedenko theorem. For a predetermined threshold u, consider the exceedance over u, denoted by $Y = X - u$. If the exceedance Y is positive (or $X > u$) for large enough u, then the conditional distribution function of Y can be approximated by the generalized Pareto (GP) distribution function

$$H(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma^*}\right)^{-1/\xi}, & \text{if } \xi \neq 0; \\ 1 - \exp\left(-\frac{y}{\sigma^*}\right), & \text{if } \xi = 0, \end{cases}$$

where $y > 0$, $1 + \xi y/\sigma^* > 0$, and $\sigma^* = \sigma + \xi(u - \mu)$. The selection of threshold u is often arguable. If u is too small, the GP distribution would be no longer appropriate for the threshold exceedances. However, if u is too large, there are not many threshold exceedances for model fit. Mean residual life (MRL) plots and parameter stability (PS) plots are often used in practice to visually find an adequate threshold (see Coles 2001).

Threshold exceedances methods are different from block maxima methods in that threshold methods use all observations above a certain threshold, instead of using only the maximum statistics within preset blocks. To elaborate, suppose n data values are greater than a threshold u and denote them as $X_1^*, \ldots, X_n^*$ (that is, $X_t^* > u$). Then consider the threshold exceedances $Y_t = X_t^* - u$. Extreme value theory implies that, if u is large enough, the GP distribution adequately approximates the probability distribution for the exceedances $\{Y_1, \ldots, Y_n\}$ regardless of $F(\cdot)$ from which the original sample $\{X_1, \ldots, X_N\}$ is taken. Since a larger number of extreme data is often used, threshold exceedances methods can perform better than block maxima methods. This is very important for extreme value studies due to the limited frequency of extreme events in many practices. The two GP parameters, σ^* and ξ, are often estimated by the maximum likelihood method.

Denote d as the number of observations per year and $p_u^* = P(X > u)$ as the over-threshold probability for a threshold u. If K is large enough such that $x_K > u$, then the return level x_K is

$$x_K = \begin{cases} u + \frac{\sigma^*}{\xi}[(Kdp_u^*)^{\xi} - 1], & \text{if } \xi \neq 0; \\ u + \sigma^*\ln(Kdp_u^*), & \text{if } \xi = 0. \end{cases} \quad (2)$$

The probability p_u^* can be estimated by an empirical probability $\hat{p}_u^* = n/N$.

3.3. Bias Correction for Temporal Dependence

As the asymptotic results in Section 3.1 and 3.2 are supported under the assumption that $X_1, \ldots, X_N$ are IID, temporal correlation in the data can induce an estimation bias for important extreme characteristics. Specifically, return level estimates could be severely biased if the data are dependent. To correct this bias caused by temporal dependence, past authors have incorporated an additional parameter, called extremal index. Suppose $X_1, \ldots, X_k$ are IID with a common marginal distribution function $F(\cdot)$ and $\tilde{X}_1, \ldots, \tilde{X}_k$ are stationary with the same distribution function $F(\cdot)$. Define $M^{(k)} = \max\{X_1, \ldots, X_k\}$ and $\tilde{M}^{(k)} = \max\{\tilde{X}_1, \ldots, \tilde{X}_k\}$. Under the assumption that the stationary series $\{\tilde{X}_1, \ldots, \tilde{X}_k\}$ satisfies the distributional mixing $D(u_k)$ condition (cf. Leadbetter, Lindgren, and Rootzen 1983, 52–54),

$$P\left(\frac{M^{(k)} - a_k}{b_k} \leq x\right) \longrightarrow G(x) \quad \text{as } k \to \infty$$

for sequences of constants $\{a_k\}$ and $\{b_k > 0\}$, if and only if

$$P\left(\frac{\tilde{M}^{(k)} - a_k}{b_k} \leq x\right) \longrightarrow G^{\theta}(x) \quad \text{as } k \to \infty$$

for a constant $0 < \theta \leq 1$. For an increasing sequence of u_k with increasing k, the $D(u_k)$ condition ensures that for any two sets of variables that are far enough apart, the two sets of variables are nearly independent, resulting in no effect on the limit laws for extremes (see Coles 2001). The extremal index θ is equal to 1 for a perfectly independent series, and less than 1 for a dependent series. Exceedances over an increasing threshold from a dependent series tend to appear in clusters in the limit. The extremal index θ is subsequently related to the sizes of these clusters, approximately equal to the reciprocal of the limiting mean cluster size (Leadbetter 1983).

To estimate the extremal index, although we do not particularly favor one single estimation method, we use the intervals estimator of Ferro and Segers (2003). This estimator does not need a specific form of dependence structure in extreme data, performing well compared to other estimators (Fawcett and Walshaw 2012). To elaborate, define $\tau_1 < \cdots < \tau_n$ as the times when the threshold exceedances $Y_1, \ldots, Y_n$ are observed. Define $\Delta_t = \tau_{t+1} - \tau_t$ as the interexceedance time for $t = 1, \ldots, n-1$. Ferro and Segers (2003) derived the limiting distribution of the interexceedance times from a strictly stationary sequence of random variables and used a moment-based approach to obtain the following estimator of θ:

$$\hat{\theta} = \min\left\{\frac{2[\sum_{t=1}^{n-1}(\Delta_t - a_1)]^2}{(n-1)\sum_{t=1}^{n-1}(\Delta_t - a_1)(\Delta_t - a_2)}, 1\right\},$$

where $a_1 = a_2 = 0$ if $\max\{\Delta_1, \ldots, \Delta_{n-1}\} \leq 2$, and $a_1 = 1$ and $a_2 = 2$ otherwise.

Ferro and Segers (2003) also proposed a bootstrap technique to construct an approximate distribution of the intervals estimator $\hat{\theta}$. Since θ^{-1} represents the limiting mean cluster size, one can assume that there are approximately independent $L = \lceil \theta n \rceil$ clusters in the threshold exceedances $Y_1, \ldots, Y_n$. The bootstrap technique then characterizes interexceedance times $\Delta_1, \ldots, \Delta_{n-1}$ into two categories: intercluster times and intracluster times. The intercluster times are defined as the largest $L-1$ interexceedance times, denoted by $\Delta_{(1)}, \ldots, \Delta_{(L-1)}$. In case of $\Delta_{(L-1)} = \Delta_{(L)}$, reduce L until $\Delta_{(L-1)}$ is strictly greater than $\Delta_{(L)}$. These intercluster times are chronologically reordered, denoted by $\Delta_{c_1}, \ldots, \Delta_{c_{L-1}}$ with $c_1 < \cdots < c_{L-1}$, exclusively partitioning the remaining interexceedance times to approximately independent L sets of intracluster times. To elaborate, let $\mathcal{A}_j = \{\Delta_{c_{j-1}+1}, \ldots, \Delta_{c_j-1}\}$ with $c_0 = 0$ and $c_L = n$ for $j = 1, \ldots, L$. If $c_j = c_{j-1} + 1$, define $\mathcal{A}_j = \{\,\}$. Then, $\{\mathcal{A}_1, \ldots, \mathcal{A}_L\}$ is a collection of approximately independent sets of intracluster times, with each $\mathcal{A}_j$ separated by the next following intercluster time Δ_{c_j}. The proposed bootstrap technique randomly chooses a sample of $L-1$ intercluster times from $\{\Delta_{c_1}, \ldots, \Delta_{c_{L-1}}\}$ and subsequently selects L sets of intracluster times from $\{\mathcal{A}_1, \ldots, \mathcal{A}_L\}$. Next, intercalate these sampled intercluster and intracluster times to form a bootstrap sample of interexceedance times. The estimated extremal index $\hat{\theta}^{(b)}$ is then calculated from each bootstrap sample for $b = 1, \ldots, B$.

For large enough K such that $x_K^* > u$, the return level x_K^* for a dependent series is

$$x_K^* = \begin{cases} u + \frac{\sigma^*}{\xi}[(Kdp_u^*\theta)^{\xi} - 1], & \text{if } \xi \neq 0; \\ u + \sigma^* \ln(Kdp_u^*\theta), & \text{if } \xi = 0. \end{cases} \quad (3)$$

Note that $x_K^* = x_K$ in (2) when $\theta = 1$. The maximum likelihood estimate for x_K^* is obtained by using the associated GP parameter estimates into (3).

3.4. Uncertainty Correction for Spatial and Temporal Dependence

Snow data recorded from one station are spatially correlated with data from its neighboring stations. In practice, the maximum likelihood estimation often uses a log-likelihood function constructed under the assumption that the data are IID and finds the values of the unknown parameters by equating the first partial derivatives of log-likelihood function to zero. Although this assumption could not be perfectly met in real-world data, a common practice is to use these score function equations as a set of the estimating equations for the unknown parameters. Using this approach, Smith (1990) found the maximum likelihood estimates under the IID assumption and then adjusts the standard error of the maximum likelihood estimates for spatial dependence. Technical details of this method are illustrated in the appendix. We propose extending this approach to our data with spatial and temporal dependence.

Temporal dependence in extreme values can be a negligible issue for block maxima methods when block size is large enough. Since our block maxima method is applied to annual maximum snowfall series, the autocorrelations between two consecutive years are in fact very low (-0.041 in Central Park, 0.014 in Newark, -0.045 in La Guardia, and -0.188 in JFK). In contrast, as threshold exceedances can be significantly autocorrelated for a short period of time, threshold exceedances methods need to take into account temporal dependence to obtain accurate results.

To address temporal dependence in threshold exceedances, decorrelation techniques were developed. These techniques aim to extract a set of independent values from all threshold exceedances. Past authors often applied the *runs-declustering* algorithm that considers all exceedances within a user-specific parameter to be a cluster and then uses only the maximum value of all exceedances within each cluster (cf. Davison and Smith 1990). However, Fawcett and Walshaw (2007) showed that the declustering method can systemically incur bias in model parameter estimates, producing underestimated return levels when threshold exceedances are strongly autocorrelated. They instead used *all* threshold exceedances and proposed modifying the standard error of the maximum likelihood estimates for autocorrelation in the threshold exceedances by incorporating Smith's method. Fawcett and Walshaw (2012) showed that using all threshold exceedances with Smith's adjustment method performs better than the declustering method for return level estimation, provided that the extremal index is appropriately estimated.

Extremes are rare by their nature. With the temporal correlation in threshold exceedances taken into account, we can further improve estimation accuracy by incorporating all neighboring, spatially correlated stations into a threshold exceedance model. Thus, this approach can use the maximum possible number of extreme observations available to us. We propose extending Smith's method to this spatially and temporally dependent extreme data, as Smith's method has been shown to perform well under spatial dependence and under temporal dependence. In our study, the log-likelihood function for a p-dimensional vector of unknown model parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, from all stations' data is reexpressed as

$$\ell_m(\boldsymbol{\theta}) = \sum_{i=1}^{m} h_i(\boldsymbol{\theta}) = \sum_{i=1}^{m}\sum_{s=1}^{g}\sum_{t=1}^{n_{is}} h_{ist}(\boldsymbol{\theta}), \quad (4)$$

where h_{ist} is a contribution from the tth observation at station s in ith year to ℓ_m, g is the number of weather stations, and n_{is} is the number of observations in station s for ith year. We assume

that the yearly contribution h_i in (4), which is constructed from observations with some dependence in a year, is independent of each other. If the data are spatially and temporally dependent but independent in each year, we can construct h_i in this manner.

3.5. Confidence Intervals for Return Levels

The delta method is often used to construct asymptotic confidence intervals for a function of parameters. However, the asymptotic confidence intervals based on the delta method can inadequately capture the sampling variability of the maximum likelihood estimators for return levels under the GEV or GP distribution settings, often failing to reach intended coverage probabilities (see Rust et al. 2011).

We consider bootstrap confidence intervals for return levels. To obtain a bootstrap confidence interval, one independently selects a random sample of size n with replacement from observed data for B times. However, this typical bootstrap method often fails to approximate the true distribution of the estimators of interest when the data are not IID (cf. Givens and Hoeting 2013). There are many techniques available to address this problem. We choose the moving block bootstrap, first introduced by Künsch (1989), due to its effectiveness and simplicity. To implement the moving block bootstrap, $n - l + 1$ overlapping blocks, each with length l, are first defined from observed data. Then, n/l blocks are randomly chosen from these blocks to obtain a bootstrap sample. For example, if we have the data $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$ with sample size $n = 9$ and choose $l = 3$, a possible moving block bootstrap sample is $\{x_3, x_4, x_5, x_7, x_8, x_9, x_1, x_2, x_3\}$. From the bth bootstrap sample, the maximum likelihood estimates of model parameters are found to compute the return level estimate $\hat{x}_K^{(b)}$ for $b = 1, \ldots, B$.

To construct a bootstrap confidence interval from the return level bootstrap estimates $\{\hat{x}_K^{(1)}, \ldots, \hat{x}_K^{(B)}\}$, one would consider the classical percentile interval method that uses the $\alpha/2$th lower and upper quantiles of the return level bootstrap estimates as confidence intervals' endpoints. However, this percentile bootstrap method often results in biased confidence intervals if the bootstrap distribution is skewed. Efron (1987) developed the bias-corrected and accelerated (BCa) bootstrap method to correct bias due to skewness. The BCa method first computes the bias-correction constant z_{BC} by

$$z_{\mathrm{BC}} = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I(\hat{x}_K^{(b)} < \hat{x}_K)\right),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, $I(E)$ is an indicator function which returns 1 if E is true and 0 otherwise, and $\hat{x}_K$ is the estimated return level from the original data. Next, the acceleration constant c_{A} is computed by

$$c_{\mathrm{A}} = \frac{\sum_{t=1}^{n}(\ddot{x}_K^{(-t)} - \ddot{x}_K)^3}{6\left[\sum_{t=1}^{n}(\ddot{x}_K^{(-t)} - \ddot{x}_K)^2\right]^{3/2}},$$

where $\ddot{x}_K^{(-t)}$ is the delete-1 jackknife estimate of x_K with tth observation deleted from the data, and $\ddot{x}_K = \frac{1}{n}\sum_{t=1}^{n}\ddot{x}_K^{(-t)}$ (cf. Givens and Hoeting 2013). Then, the $(1-\alpha)\times 100\%$ BCa interval uses the following quartiles of the return level bootstrap estimates as the endpoints:

$$\Phi\left(z_{\mathrm{BC}} + \frac{z_{\mathrm{BC}} - z_{\alpha/2}}{1 - c_{\mathrm{A}}(z_{\mathrm{BC}} - z_{\alpha/2})}\right),$$
$$\Phi\left(z_{\mathrm{BC}} + \frac{z_{\mathrm{BC}} + z_{\alpha/2}}{1 - c_{\mathrm{A}}(z_{\mathrm{BC}} + z_{\alpha/2})}\right),$$

where $z_{\alpha/2}$ is the $\alpha/2$th upper quantile of the standard normal distribution. We use this method to obtain confidence intervals for return levels.

4. Results

4.1. Snowfall GEV Models With Return Levels

We apply the block maxima method to the annual maximum snowfall series at the Central Park, Newark, La Guardia, and JFK stations. For this, suppose that N_s snowstorms occurred at station s during the study period of 1959–2014. Let $\{X_{s1}, \ldots, X_{sN_s}\}$ denote the snowfall series at station s, where X_{sj} represents the accumulated snowfall corresponding to the jth snowstorm for the data period. Denote M_{st} as the maximum snowfall observation at station s in snow year (block) $t = 1959, \ldots, 2014$. To account for possible geographical location effects on snowfall, we assume that $\{M_{st}\}$ follows a GEV(μ_s, σ_s, ξ_s) distribution with μ_s as the location parameter for each of the four stations. For model selection, we use the corrected Akaike information criterion: AICc $= 2p - 2\ell + 2p(p+1)/(n-p-1)$, where ℓ is the maximum log-likelihood value for model, p is the number of model parameters, and n is the sample size. The model with the lowest AICc value is preferred. We find that a GEV model with different σ_s's and ξ_s's does not improve a GEV model fit. Specifically, a model with different μ_s's, σ_s's, and ξ_s's returned the AICc value of 1376.398, whereas a model with different μ_s's but σ and ξ fixed returned a more appealing AICc value of 1362.943. As the four stations are geographically close, this result is expected. Hereafter, we consider only those models with different μ_s's but σ and ξ fixed for all four stations.

GEV estimation results are summarized in Table 2. We considered the four different stationary GEV models with varying μ_s's by iteratively merging the two closest μ_s's. We then compute the standard errors of GEV maximum likelihood estimates. The naïve standard errors (ignoring spatial dependence among the four stations) and the corrected standard errors by Smith's method (including spatial correlation for the stations) are displayed in the table next to the parameter estimates. A sample of spatially correlated data could provide less information than a sample of independent data with the same sample size, adding more uncertainty to the GEV parameter estimates. Due to this, the spatial correlation corrected standard errors are greater than the corresponding naïve standard errors. We choose Model 3 for further return level analysis as this model produces the lowest AICc value.

Annual maximum snowfall data could show nonstationary characteristics, such as long-term trends. Specifically, to quantify a possible long-term linear trend in the annual maximum snowfall series at station s, we reparameterize the GEV location parameter as

$$\mu_{st} = \mu_s + \beta\left(\frac{t - 1958}{10}\right)$$

Table 2. Stationary GEV estimates for annual maximum snowfall data with their associated standard errors in parentheses (*left*: naïve, *right*: corrected).

		Model 1	Model 2	Model 3	Model 4
μ	μ_1	6.952 (0.484, 0.545)	6.863 (0.375, 0.589)	6.748 (0.322, 0.577)	6.556 (0.293, 0.571)
	μ_2	6.760 (0.512, 0.658)			
	μ_3	6.526 (0.491, 0.578)	6.535 (0.491, 0.573)		
	μ_4	5.981 (0.488, 0.578)	5.988 (0.489, 0.574)	5.990 (0.489, 0.573)	
σ		3.788 (0.229, 0.482)	3.793 (0.228, 0.481)	3.797 (0.228, 0.480)	3.823 (0.230, 0.473)
ξ		0.168 (0.059, 0.085)	0.165 (0.058, 0.084)	0.165 (0.058, 0.084)	0.162 (0.059, 0.081)
ℓ		−674.212	−674.254	−674.426	−675.487
AICc		1362.943	1360.895	1359.127	1359.157

Table 3. Nonstationary GEV estimates for annual maximum snowfall data with their associated standard errors in parentheses (*left*: naïve, *right*: corrected).

		Model 1	Model 2	Model 3	Model 4
μ	μ_1	6.281 (0.656, 0.911)	6.214 (0.562, 0.950)	6.114 (0.523, 0.929)	5.962 (0.514, 0.927)
	μ_2	6.181 (0.656, 1.010)			
	μ_3	5.953 (0.639, 0.911)	5.918 (0.632, 0.905)		
	μ_4	5.287 (0.677, 0.930)	5.288 (0.671, 0.926)	5.293 (0.671, 0.923)	
	β	0.246 (0.154, 0.298)	0.231 (0.152, 0.293)	0.232 (0.152, 0.292)	0.213 (0.153, 0.297)
σ		3.796 (0.233, 0.473)	3.768 (0.229, 0.459)	3.772 (0.230, 0.458)	3.807 (0.230, 0.454)
ξ		0.156 (0.061, 0.083)	0.169 (0.061, 0.081)	0.169 (0.061, 0.080)	0.162 (0.060, 0.077)
ℓ		−673.136	−673.121	−673.263	−674.523
AICc		1360.791	1358.629	1356.801	1357.229

for snow year $t = 1959, \ldots, 2014$. Here, the trend parameter β is interpreted as the expected change in maximum snowfall over a decade, since

$$E(M_{s,t+10}) - E(M_{s,t}) = \left[\mu_{s,t+10} + \tfrac{\sigma}{\xi}(\Gamma(1-\xi)-1)\right] - \left[\mu_{s,t} + \tfrac{\sigma}{\xi}(\Gamma(1-\xi)-1)\right] = \beta$$

when $\xi < 1$. We assume that the geospatial trends in maximum snowfall series are the same for the four stations.

The nonstationary GEV estimation results are summarized in Table 3. We select Model 3 for further return level analysis as this model is parsimonious and a good fit for the data according to the AICc values. The estimated linear trend in Model 3 is 0.232 inches decade^{-1}. Based on 56 annual maximum snowfall observations from each of the four stations, the trend estimate β is not significantly greater than zero with the corrected standard error 0.292 used for significance test. However, we note that longer time series may provide a clearer picture by reducing estimation uncertainty associated with trends in the maximum snowfall.

The return levels for 25, 50, 75, and 100 years based on the selected Model 3 stationary and nonstationary GEV models are estimated using the expression (1) with the corresponding maximum likelihood estimates substituted. For interval estimation of return levels, the typical confidence interval based on asymptotic theory, expressed as an estimate ± (1.96×standard error), can be considered. However, as the shape parameter estimate is positive, the GEV distribution is right skewed, implicating that the return level distribution is also right skewed. We use bootstrap confidence intervals to approximate this right skewness. To further correct bias in conventional percentile bootstrap methods, we compute the BCa bootstrap confidence intervals, as illustrated in Section 3.5, with 10,000 replications.

Figure 1 displays the resulting 95% BCa bootstrap intervals. Although the maximum likelihood estimates of return levels calculated with and without trends are different, it is worth mentioning that the associated 95% confidence intervals are not drastically different. This further verifies that there are no significant long term linear trends in the annual maximum snowfall series at this point. The actual snowfall brought by the January 2016 snowstorm is about equivalent to a 50-year return level for Central Park, Newark, and La Guardia; and 75-year return level for JFK when linear trends are not considered. However, if linear trends are included, the actual snowfall is between 25 and 50-year return levels for Central Park, Newark, and La Guardia; and between 50- and 75-year return levels for JFK.

4.2. Snowfall GP Models with Return Levels

Due to a limited number of extreme events, threshold exceedances methods can provide additional meaningful information in understanding extreme events. To determine an adequate threshold u, we used two graphical methods: MRL plots and PS plots. In Figure 2, the MRL plot suggests that a threshold of 87.5th percentile of all nonzero snowfall observations (around 6 inches) could be adequate. This threshold candidate can be also justified on practical viewpoints because the National Weather Service Forecast Office in New York City issues a winter storm warning when 6 inches or more of snow is expected in a 12-hour period or at least 8 inches in a 24-hour period. The PS plot in Figure 2 also supports this threshold candidate. Therefore, we choose thresholds for nonzero snowfall series to be 87.5th percentile so that about 12.5% of observations would exceed this threshold.

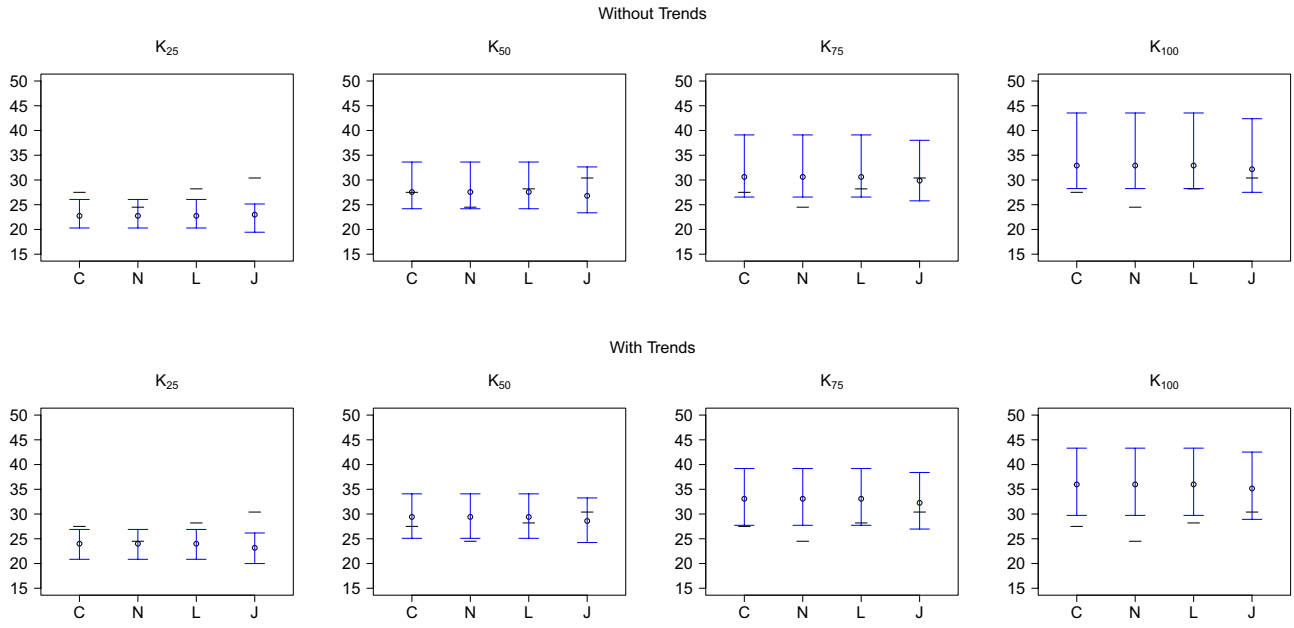

Figure 1. GEV annual maximum snowfall return level estimates (○), their associated 95% BCa bootstrap confidence intervals (⌶), and actual 2016 blizzard snowfall (—).

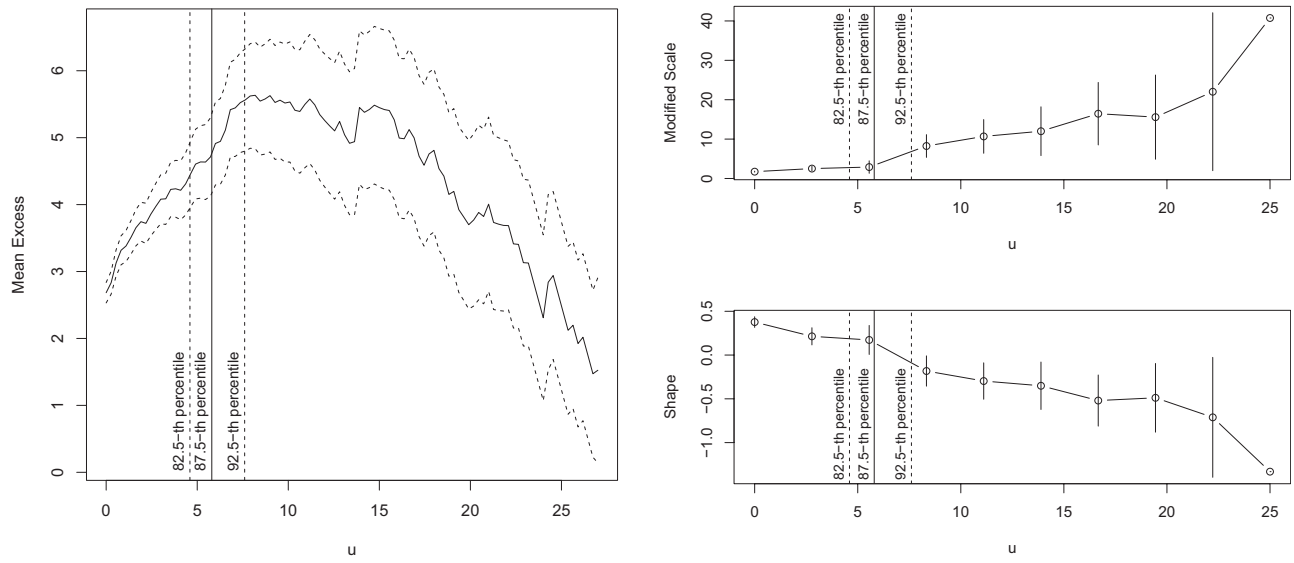

Figure 2. (*Left*) Mean residual life plot for all nonzero snowfall observations and (*Right*) parameter stability plots for GP fit to all nonzero snowfall observations (*u*: threshold).

We then fit a GP distribution to all exceedances over the 87.5th percentile threshold. For this, assume that u_s is the predetermined threshold associated with station s. Let $\{X_{s1}, \ldots, X_{sN_s}\}$ denote the snowfall series at station s for the study period, where X_{sj} represents the accumulated snowfall corresponding to the jth snowstorm. If $X_{sj} > u_s$, we denote such X_{sj} by X^*_{st}, where t refers to the day in which the corresponding snowstorm occurred. Specifically, if a snowstorm lasted for one day, t denotes the corresponding date of occurrence in $\{1, \ldots, 20454\}$, with day 1 as July 1, 1959 (the first day of snow year 1959), and day 20,454 as June 30, 2015 (the last day of snow year 2014 and last day of the study period). If a snowstorm had lasted for two or more days, the average of corresponding dates of occurrence is used as the value for t. The over-threshold exceedance is then $Y_{st} = X^*_{st} - u_s$. We assume that these positive Y_{st}'s have a GP$(\sigma^*_s, \xi_s, \theta_s)$ distribution for each station s. As discussed in Section 4.1, we assume that σ^*_s's and ξ_s's are identical within each s. In fact, our GP estimation result for different scale and shape parameters did not produce a significantly improved model. Also, our estimates of ξ_s's are not significantly different from zero for all cases (Models 1, 2, 3, and 4), confining to the GP models with identical $\sigma^*_s = \sigma^*$ and $\xi_s = 0$. The extremal indexes θ_s's are estimated using the intervals estimator explained in Section 3.3.

With the initial thresholds selected by MRL and PS plots, our threshold exceedances method proceeds by following two steps. First, a quantile regression model is applied to obtain an optimal threshold. Second, GP maximum likelihood estimation is performed for over-threshold exceedances. Quantile regression methods estimate the conditional quantiles for the probability distribution of a response variable, producing a fitted curve that corresponds to various percentage points for the response

variable (see Koenker 2005). We obtain stationary thresholds by fitting a 87.5th quantile regression with different intercepts u_s's from all nonzero snowfall observations at each station. For a more parsimonious model, the two closest threshold u_s's are merged in the quantile regression step. The corresponding GP σ_s^*'s are estimated under the assumption $\xi_s = 0$ in the next GP estimation step. This two-step procedure is repeated until no other two u_s's are noticeably different, with θ_s's also merged along with the u_s's.

Table 4 summarizes our two-step estimation results. The standard errors for quantile regression parameters are computed using the *se=boot* option within the *summary.rq* function in the R package *quantreg*, which implements an effective bootstrap technique (Koenker 2016). The default *xy-pair* method is used. The standard errors for the maximum likelihood estimate of GP scale parameter σ^* are computed using two different approaches. The first approach is the naïve method that ignores spatial and temporal dependence in observed data. The second approach is our corrected method that extends Smith's method to account for both spatial and temporal dependence. The corrected method inflates the GP standard errors by a factor of two, which is similar to our finding in the GEV cases. This implies that spatial and temporal dependence contributes to larger uncertainty in GP parameter estimates. The standard errors for extremal indexes are computed using the bootstrap method proposed by Ferro and Segers (2003) as illustrated in Section 3.3.

Under climate change scenarios, use of constant thresholds can cause inaccurate results. For example, a constant threshold that is adequate for the observations in the period of 1960–1980 could be different from those in the period of 1990–2010. Recent studies (Kyselỳ, Picek, and Beranová 2010; Northrop and Jonathan 2011; Jonathan, Ewans, and Randell 2014) used quantile regression methods to obtain time-dependent thresholds for threshold modeling. Therefore, we consider non-stationary thresholds to quantify a possible long-term linear trend in the over-threshold snowfall exceedances. Assuming that the trends in nonzero snowfall series are identical for the four stations, we estimate the time-dependent threshold u_{st} by fitting a 87.5th quantile regression model with different intercepts u_s's and a linear trend:

$$u_{st} = u_s + \beta\left(\frac{t}{365.25 \times 10}\right).$$

This time-dependent threshold u_{st} can more likely keep the over-threshold probability p_u^* constant over the entire study period.

Table 5 summarizes our quantile regression and GP maximum likelihood estimation results. The estimated linear trend in Model 3 is 0.564 inches decade^{-1}, which is about 2.4 times greater than the GEV estimated trend for annual maximum snowfall series. In addition, the standard error for Model 3 with all exceedances is much smaller than the standard error for the GEV trend estimate with annual maximum series. This is partly

Table 4. Stationary GP estimates and their associated standard errors in parentheses (*left*: naïve, *right*: corrected) with the assumption $\xi = 0$.

		Model 1	Model 2	Model 3	Model 4
u	u_1	6.000 (0.368)	6.000 (0.368)	5.900 (0.217)	5.800 (0.181)
	u_2	5.700 (0.443)	5.800 (0.323)		
	u_3	5.900 (0.425)			
	u_4	5.500 (0.453)	5.500 (0.453)	5.500 (0.453)	
σ^*		4.652 (0.280, 0.493)	4.601 (0.275, 0.491)	4.555 (0.272, 0.487)	4.542 (0.270, 0.486)
θ	θ_1	0.838 (0.089)	0.838 (0.089)	0.806 (0.090)	0.891 (0.054)
	θ_2	0.841 (0.089)	0.880 (0.076)		
	θ_3	0.736 (0.114)			
	θ_4	0.926 (0.074)	0.926 (0.074)	0.926 (0.074)	
ℓ		−702.831	−707.339	−707.047	−708.760

Table 5. Nonstationary GP estimates and their associated standard errors in parentheses (*left*: naïve, *right*: corrected) with the assumption $\xi = 0$.

		Model 1	Model 2	Model 3	Model 4
u	u_1	4.652 (0.559)	4.644 (0.600)	4.633 (0.607)	4.109 (0.363)
	u_2	4.044 (0.487)	4.112 (0.395)	3.986 (0.381)	
	u_3	4.232 (0.490)			
	u_4	3.919 (0.512)	3.850 (0.455)		
	β	0.525 (0.122)	0.541 (0.110)	0.564 (0.116)	0.548 (0.111)
σ^*		4.582 (0.272, 0.500)	4.529 (0.268, 0.499)	4.527 (0.269, 0.502)	4.618 (0.275, 0.513)
θ	θ_1	0.879 (0.087)	0.879 (0.087)	0.892 (0.083)	0.895 (0.059)
	θ_2	0.924 (0.077)	0.899 (0.077)	0.901 (0.066)	
	θ_3	0.877 (0.096)			
	θ_4	0.952 (0.085)	0.998 (0.065)		
ℓ		−713.739	−717.976	−712.870	−716.002

due to the fact that the quantile regression and GP model use all non-zero observations, therefore contributing to the reduction of variance. The estimated trend in all exceedances is determined to be statistically significant: 87.5th percentile snowfall has increased by 0.564 inches decade^{-1}. This is in contrast to the annual maximum snowfall, which shows an insignificant increase by 0.232 inches decade^{-1}.

To explore different effects at different percentiles, we also consider two additional thresholds: 82.5th and 92.5th percentiles. The MRL and PS plots in Figure 2 suggest that these threshold candidates are also acceptable. Table 6 summarizes our additional results from the quantile regression and GP maximum likelihood estimation. The quantile regression trend estimates are statistically significant, implying 82.5th percentile and 92.5th percentile snowfalls have increased by 0.495 and 0.661 inches decade^{-1}, respectively. The extremal index estimates are in the range of 0.8 and 0.9.

We now perform return level analysis based on the nonstationary GP Model 3 as this model has shown to offer significantly improved fitting over the other GP models. As we combined consecutive nonzero daily snowfall observations during a single snowstorm event period into one snowfall observation for the snowstorm event period, each snowfall observation in our data is no longer equally weighted. Consequently, an adjustment is needed to calculate the over-threshold probability $p_u^* = P(X > u)$. We estimate p_u^* in (3) by n^*/N^*, where n^* is the total number of days corresponding to each threshold-exceeding snowfall observation and N^* is the total number of days corresponding to all snowfall observations. For the value of d in (3), we use the average of 56 yearly snowfall observation counts.

Figure 3 shows the estimated return levels for 25, 50, 75, and 100 years from the nonstationary GP Model 3 along with their 95% confidence intervals using the moving block BCa bootstrap method with 10,000 replications. The return level estimates computed from nonstationary GP models align fairly well with those from GEV models. The estimates from the 82.5th percentile model are closest to the estimates from the GEV models. Overall, the confidence intervals in Figure 3 are narrower than those in Figure 1, with the 82.5th percentile model producing much narrower return level intervals than the 92.5th percentile model. These results implicate that more observations above the thresholds contribute to the reduction of uncertainty for GP models. The actual snowfall from the 2016 blizzard are approximately equivalent to a 25-year return level for Central Park and La Guardia, less than a 25-year return level for Newark, and a 50-year return level for JFK.

5. Comments and Conclusions

Extreme value analysis requires use of an appropriate method based on extreme value theory. Using nonextreme value distributions to estimate return levels of extreme data could result in tremendous bias. To elaborate, we fit a gamma distribution model to the NRCC's snow data and assess the practicability of the gamma distribution model in terms of return periods, which are an inverse of return levels. This nonextreme value distribution model has been frequently used in the literature. Table 7 summarizes the estimated shape and rate parameters of the gamma distribution model and associated return periods in years for the 2016 blizzard. The gamma model produces unrealistically high return periods for the actual 2016 blizzard observations. Therefore, other models should be used to explain extreme characteristics of the snow data.

We comment on return periods for the 2016 blizzard using extreme value theory. As explained in Section 4, our return periods are calculated based on the three extreme value methods: 1. stationary GEV block maxima method, 2. non-stationary GEV block maxima method with linear trend, and 3. nonstationary GP threshold exceedances method with extremal index and linear trend. Table 8 summarizes our estimated return periods from the two GEV methods and the nonstationary GP method using the 87.5th percentile threshold. The results show more realistic return period estimates for the 2016 blizzard than those from the gamma distribution model. For the period of January 22–24, 2016, JFK had a higher snowfall record (30.4 inches) than the other three stations (27.5, 24.5, and 28.2 inches), resulting in JFK having a return period of 80, 56, and 45 years for the three methods.

We now include the snow year 2015 (July 1, 2015–June 30, 2016) data in the three extreme value methods to see how much the model parameters and return levels change. For the station-

Table 6. Nonstationary GP Model 3 estimates using three thresholds and the associated standard errors in parentheses (*left*: naïve, *right*: corrected) with the assumption $\xi = 0$.

		82.5th percentile	87.5th percentile	92.5th percentile
u	u_1	3.373 (0.272)	4.633 (0.607)	6.393 (0.499)
	u_2		3.986 (0.381)	
	u_3			5.464 (0.536)
	u_4	2.947 (0.341)		
	β	0.495 (0.105)	0.564 (0.116)	0.661 (0.204)
σ^*		4.122 (0.207, 0.413)	4.527 (0.269, 0.502)	4.985 (0.385, 0.605)
θ	θ_1	0.838 (0.093)	0.892 (0.083)	0.807 (0.115)
	θ_2		0.901 (0.066)	
	θ_3			0.827 (0.070)
	θ_4	0.796 (0.060)		
ℓ		−959.331	−712.870	−437.896

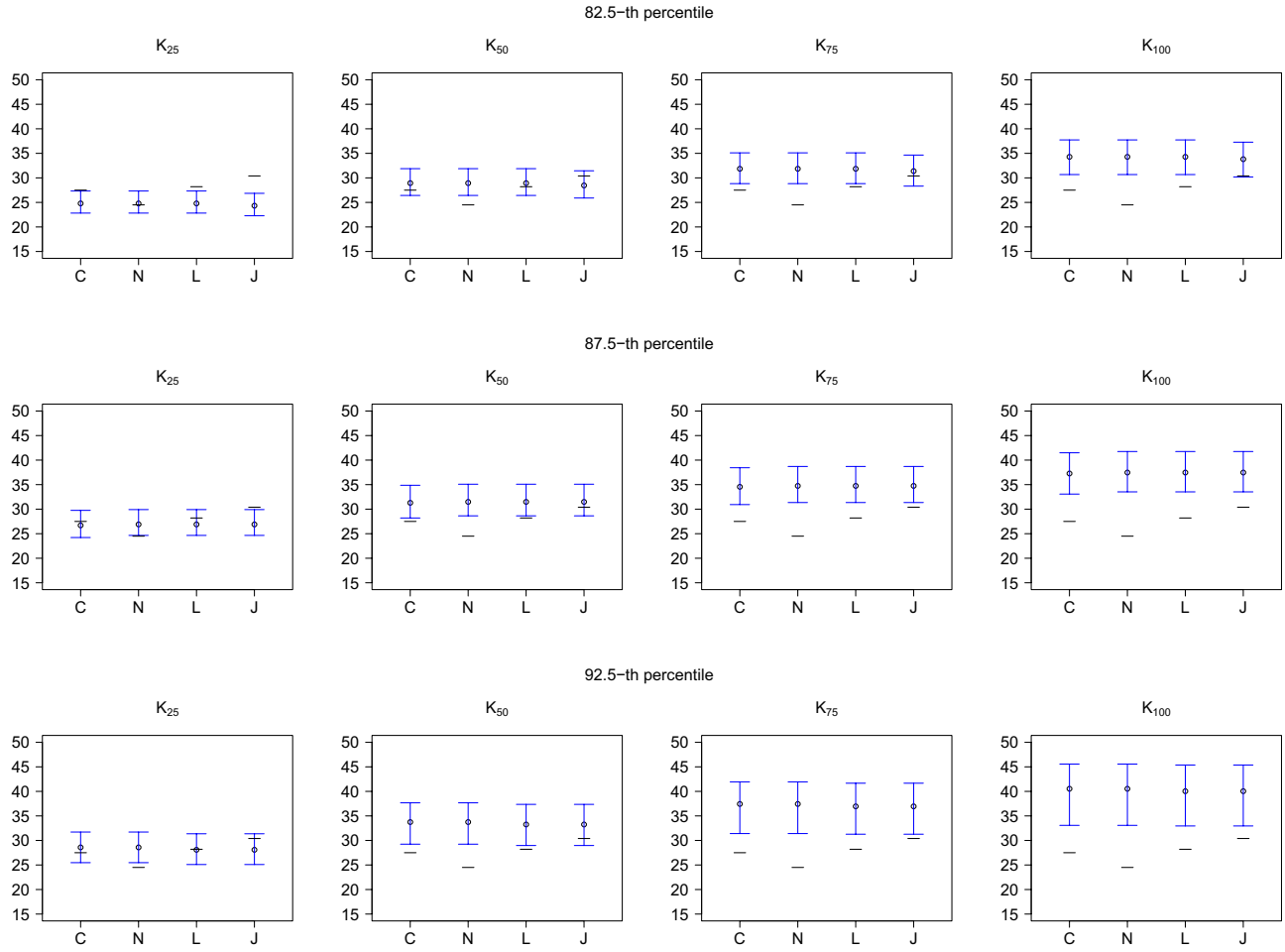

Figure 3. Nonstationary GP snowfall return level estimates (○), their associated 95% BCa bootstrap confidence intervals (⌶), and actual 2016 blizzard snowfall (—).

Table 7. Gamma distribution model estimation results and their return periods.

	Shape	Rate	Return period
Central Park	0.847	0.286	444 years
Newark	0.701	0.264	135 years
La Guardia	0.788	0.302	926 years
JFK	0.780	0.320	3547 years

Table 8. Return periods for the actual snowfall observations from the January 2016 blizzard.

	Actual	GEV return period		GP return period
	observation (inches)	Stationary (years)	Nonstationary (years)	Nonstationary (years)
Central Park	27.5	50	40	26
Newark	24.5	32	27	18
La Guardia	28.2	55	43	32
JFK	30.4	80	56	45

ary and nonstationary GEV block maxima methods, including the maximum snowfall observations from the snow year 2015 results in about a 19% increase of the maximum likelihood estimate of shape parameter. The nonstationary GEV method produces the estimated snowfall linear trend to increase by 17%, although the trend estimate is still not significantly greater than zero with the corrected standard errors. These changes in the GEV parameters further result in an increase of up to about three inches in return levels for the four stations in the New York City area. For the nonstationary GP threshold exceedances method, the maximum likelihood estimates of scale and linear trend parameters show little to no increases after including the snow year 2015 nonzero snowfall observations. However, these changes in GP parameters result in an increase up to about one inch to return levels for the four stations.

We found insignificant upward trends for the annual maximum snowfall series in the New York City area. It would be possible to further reduce estimation uncertainty in annual maximum snowfall trends when more data become available in the future. Even with this said, we have found that the 82.5th, 87.5th, and 92.5th snowfall percentiles have significantly increased using the quantile regression and GP models. These findings indicate that the distribution of extreme snowfall in the New York City area has gradually changed over time, with more noticeable changes within high percentile snowfall events, and a possibility to observe increasing annual maximum snowfall in the future. With overall decreasing annual mean snowfall reported in the literature (see Burakowski et al. 2008) considered, significant increases in snowfall among the higher percentiles, while insignificant increases in the annual maximum snowfall, can suggest that there have been fewer snow storms with more snowfall each.

The blizzard of January 2016 in the New York City area is indeed an extreme event. We further assess this blizzard as about a once-in-forty-years snow event in terms of return periods. The extreme value analysis techniques used in this paper are adequate enough in explicating this extreme event and will help practitioners study extreme characteristics of data using extreme value theory and methods.

Appendix: Smith's Method

As Smith's method is a basis of our treatment for spatial and temporal dependence, we illustrate this method in our setting. Suppose there are multiple stations in a region and each station has m years of data recorded. Assume the log-likelihood function for a p-dimensional vector of unknown model parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, from all stations' data is expressed as

$$\ell_m(\boldsymbol{\theta}) = \sum_{i=1}^{m} h_i(\boldsymbol{\theta}), \tag{A.1}$$

where h_i is the contribution of all stations' ith year data, independent for other i's, to the log-likelihood function $\ell_m(\cdot)$. Denote $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^{\mathrm{T}}$ as the maximum likelihood estimate of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0 = (\theta_{1,0}, \ldots, \theta_{p,0})^{\mathrm{T}}$ as the true value of $\boldsymbol{\theta}$. Applying Taylor expansion to the estimating equation $\nabla \ell_m(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ produces

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx -[\nabla^2 \ell_m(\boldsymbol{\theta}_0)]^{-1} \nabla \ell_m(\boldsymbol{\theta}_0),$$

where ∇ and ∇^2 represent the following gradient and Hessian:

$$\nabla \ell_m(\boldsymbol{\theta}_0) = \begin{pmatrix} \dfrac{\partial \ell_m(\boldsymbol{\theta}_0)}{\partial \theta_{1,0}} \\ \vdots \\ \dfrac{\partial \ell_m(\boldsymbol{\theta}_0)}{\partial \theta_{p,0}} \end{pmatrix},$$

$$\nabla^2 \ell_m(\boldsymbol{\theta}_0) = \begin{pmatrix} \dfrac{\partial^2 \ell_m(\boldsymbol{\theta}_0)}{\partial \theta_{1,0}^2} & \cdots & \dfrac{\partial^2 \ell_m(\boldsymbol{\theta}_0)}{\partial \theta_{1,0} \partial \theta_{p,0}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \ell_m(\boldsymbol{\theta}_0)}{\partial \theta_{p,0} \partial \theta_{1,0}} & \cdots & \dfrac{\partial^2 \ell_m(\boldsymbol{\theta}_0)}{\partial \theta_{p,0}^2} \end{pmatrix}.$$

Approximating each entry in the Hessian by its expected value gives

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx H^{-1} \nabla \ell_m(\boldsymbol{\theta}_0), \tag{A.2}$$

where $H = -E[\nabla^2 \ell_m(\boldsymbol{\theta}_0)]$ is the Fisher information. Taking variance to both sides of (A.2) produces the variance of maximum likelihood estimators:

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}) \approx H^{-1} V H^{-1}, \tag{A.3}$$

where $V = \mathrm{cov}(\nabla \ell_m(\boldsymbol{\theta}_0))$. If the assumed model is correct (that is, if the data are spatially independent), then one obtains $V = H$, resulting in the conventional approximation $\mathrm{cov}(\hat{\boldsymbol{\theta}}) \approx H^{-1}$. The Fisher information H then can be approximated by the observed information $H \approx -\nabla^2 \ell_m(\hat{\boldsymbol{\theta}})$, which is typically calculated from software by using a quasi-Newton or similar algorithm.

Now, suppose the data are spatially dependent among stations, but the yearly contribution h_i's in (A.1) are independent. This assumption can be validated in practice. For example, consider annual maximum snowfall series observed at several stations in a study region. Maximum snowfall values are spatially correlated among the neighboring stations for a given year but are nearly uncorrelated between years at a station. Under the assumption that h_i's have an identical distribution, the covariance V in (A.3) is reexpressed as

$$\begin{aligned} V &= \mathrm{cov}\big(\nabla \ell_m(\boldsymbol{\theta}_0)\big) = \mathrm{cov}\Big(\textstyle\sum_{i=1}^{m} \nabla h_i(\boldsymbol{\theta})\Big) \\ &= \textstyle\sum_{i=1}^{m} \mathrm{cov}\big(\nabla h_i(\boldsymbol{\theta})\big) = m\mathrm{cov}\big(\nabla h_1(\boldsymbol{\theta})\big). \end{aligned}$$

Here, $\mathrm{cov}(\nabla h_1(\boldsymbol{\theta}))$ can be estimated by the empirical covariance of the observed gradient values $\nabla h_1(\hat{\boldsymbol{\theta}}), \ldots, \nabla h_m(\hat{\boldsymbol{\theta}})$. Specifically, the (j,k)th element of $\mathrm{cov}(\nabla h_1(\boldsymbol{\theta}))$ is

$$\frac{1}{m} \sum_{i=1}^{m} (u_{i,j}(\hat{\boldsymbol{\theta}}) - \bar{u}_{\cdot,j}(\hat{\boldsymbol{\theta}}))(u_{i,k}(\hat{\boldsymbol{\theta}}) - \bar{u}_{\cdot,k}(\hat{\boldsymbol{\theta}})), \quad \text{for } j,k = 1, \ldots, p,$$

where $u_{i,j}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} h_i(\boldsymbol{\theta})$ and $\bar{u}_{\cdot,j}(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} u_{i,j}(\boldsymbol{\theta})$. Using $\hat{V} = m\widehat{\mathrm{cov}}\big(\nabla h_1(\boldsymbol{\theta})\big)$ into (A.3) produces the estimated variance of maximum likelihood estimators

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\theta}}) = \hat{H}^{-1} \hat{V} \hat{H}^{-1}, \tag{A.4}$$

where $\hat{H}$ is a quasi-Newton approximation of H and is usually calculated by software under IID assumption. Smith (1990) shows that if data are spatially correlated, the variance in (A.4) is more accurate than the variance under IID assumption.

Supplementary Material

- **R code for model fitting and return level calculation**: R code for GEV and GP model fitting (only model 3) and return level calculation along with producing BCa bootstrap confidence intervals (ComputeReturnLevels.R).
- **Snowfall data**: The dataset used throughout this paper (SnowfallData.RData).

Acknowledgments

The authors thank an anonymous reviewer, associate editor, and the editor for their thoughtful comments and valuable suggestions that significantly improved this paper.

References

Blanchet, J., Marty, C., and Lehning, M. (2009), "Extreme Value Statistics of Snowfall in the Swiss Alpine Region," *Water Resources Research*, 45. W05424. [283]

Burakowski, E. A., Wake, C. P., Braswell, B., and Brown, D. P. (2008), "Trends in Wintertime Climate in the Northeastern United States: 1965–2005," *Journal of Geophysical Research*, 113. D20114. [282,291]

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*. London, UK: Springer-Verlag. [284]

Davison, A. C., and Smith, R. L. (1990), "Models for Exceedances over High Thresholds," *Journal of the Royal Statistical Society*, Series B, 52, 393–442. [285]

Efron, B. (1987), "Better Bootstrap Confidence Intervals" (with discussion), *Journal of the American Statistical Association*, 82, 171–200. [286]

Fawcett, L., and Walshaw, D. (2007), "Improved Estimation for Temporally Clustered Extremes," *Environmetrics*, 18, 173–188. [282,285]

——— (2012), "Estimating Return Levels From Serially Dependent Extremes," *Environmetrics*, 23, 272–283. [282,285]

Ferro, C. A. T., and Segers, J. (2003), " Inference for Clusters of Extreme Values," *Journal of the Royal Statistical Society,* Series B, 65, 545–556. [285,289]

Givens, G. H., and Hoeting, J. A. (2013), *Computational Statistics* (2nd ed.). Hoboken, NJ: Wiley. [286]

Huntington, T. G., Hodgkins, G. A., Keim, B. D., and Dudley, R. W. (2004), "Changes in the Proportion of Precipitation Occurring as Snow in New England (1949–2000)," *Journal of Climate*, 17, 2626–2636. [282]

Jonathan, P., Ewans, K., and Randell, D. (2014), "Non-Stationary Conditional Extremes of Northern North Sea Storm Characteristics," *Environmetrics*, 25, 172–188. [289]

Koenker, R. (2005), *Quantile Regression*, Cambridge, UK: Cambridge University Press. [289]

——— (2016), *quantreg: Quantile Regression*, R package version 5.29. [289]

Kunkel, K. E., Palecki, M., Ensor, L., Hubbard, K. G., Robinson, D., Redmond, K., and Easterling, D. (2009), "Trends in Twentieth-Century U.S. Snowfall Using a Quality-Controlled Dataset," *Journal of Atmospheric and Oceanic Technology*, 26, 33–44. [282]

Künsch, H. R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241. [286]

Kyselỳ, J., Picek, J., and Beranová, R. (2010), " Estimating Extremes in Climate Change Simulations Using the Peaks-over-threshold Method With a Non-stationary Threshold," *Global and Planetary Change*, 72, 55–68. [289]

Leadbetter, M. R. (1983), "Extremes and Local Dependence in Stationary Sequences," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65, 291–306. [284]

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983), *Extremes and Related Properties of Random Sequences and Processes* (1st ed.), New York, NY: Springer-Verlag. [284]

Lee, J., Li, S., and Lund, R. (2014), "Trends in Extreme U.S. Temperatures," *Journal of Climate*, 27, 4209–4225. [283]

López-Moreno, J. I., Goyette, S., Vicente-Serrano, S. M., and Beniston, M. (2011), " Effects of Climate Change on the Intensity and Frequency of Heavy Snowfall Events in the Pyrenees," *Climatic Change*, 105, 1–20. [283]

Makkonen, L., Ruokolainen, L., Räisänen, J., and Tikanmäki, M. (2007), "Regional Climate Model Estimates for Changes in Nordic Extreme Events," *Geophysica*, 43, 25–48. [283]

McCormick, W. P., and Qi, Y. (2000), "Asymptotic Distribution for the Sum and Maximum of Gaussian Processes," *Journal of Applied Probability*, 37, 958–971. [282]

National Weather Service (2016), *Evaluation of Reported snowfall at Local Climatological Data Stations During the East Coast Blizzard of January 22–23, 2016*, Silver Spring, MD: Office of Climate, Water and Weather Services. [282]

NOAA National Centers for Environmental Information (NCEI) (2017), *U.S. Billion-dollar Weather and Climate Disasters*, available at *http://www.ncdc.noaa.gov/billions/*. [282]

Northrop, P. J., and Jonathan, P. (2011), "Threshold Modelling of Spatially Dependent Non-stationary Extremes With Application to Hurricane-induced Wave Heights," *Environmetrics 22*, 799–809. [282,289]

O'Gorman, P. A. (2014), "Contrasting Responses of Mean and Extreme Snowfall to Climate Change," *Nature*, 512, 416–418. [282]

Panagoulia, D., Economou, P., and Caroni, C. (2014), "Stationary and Nonstationary Generalized Extreme Value Modelling of Extreme Precipitation Over a Mountainous Area Under Climate Change," *Environmetrics*, 25, 29–43. [283]

Rust, H. W., Kallache, M., Schellnhuber, H. J., and Kropp, J. P. (2011), "Confidence Intervals for Flood Return Level Estimates Assuming Long-range Dependence," in *In Extremis: Disruptive Events and Trends in Climate and Hydrology*, eds. J. Kropp and H. J. Schellnhube, Heidelberg, Germany: Springer-Verlag, pp. 60–88. [282,286]

Smith, R. L. (1990), "Regional Estimation From Spatially Dependent Data," Unpublished manuscript. [283,285,292]