Forecasting Extreme Space Weather Over Time

Isaac Finberg

Introduction

Geomagnetic storms are a result of extreme solar activity wherein severe solar winds interact with and disturb the Earth's magnetosphere. While geomagnetic storms are not generally thought of as pressing in the same manner as other climate phenomena, such as planetary warming, they can pose a threat to life on Earth when they manifest at extreme scales. Much of modern life is reliant on electronic systems, and these devices can be disrupted and damaged by geomagnetic storms. One such storm, referred to as the Quebec event, caused an electric power system failure in a large area of Canada in 1989 [1]. Another similar event in 2003 resulted in decreased GPS positioning accuracy [1]. Several indices are used to measure and track the severity of these storms, some with records spanning multiple hundreds of years. One example of these indices with easily accessible data is the Disturbance-storm time index (Dst), which measures the hourly average disturbance of the geomagnetic field in the Earth's low-latitude region. Specifically, it is composed of weighted measurements from several near-equatorial geomagnetic observatories that quantify the intensity of the globally symmetric electrojet [2].

Due to the relative rarity of events extreme enough to impact daily life, severe storms are difficult to model and predict with accuracy. A multitude of studies have attempted to model these storms using various tools, including physics-based methods, data-based models, or a combination of the two. A failure of standard statistical analyses is that the likelihood of extreme events is underestimated by the normal distribution. One area of study that is well-suited to this type of problem is extreme value theory, where distributions that characterize this field are developed to better describe the so-called heavy-tail distribution. In this work, ideas from extreme value theory and time-series analysis are employed to understand historical extreme geomagnetic behavior and attempt to estimate the return levels and periods for future severe storms.

Dataset

The dataset used in this analysis was procured from the U.S. Department of the Interior's Science Base Catalog, which houses a number of public geological datasets worked on by internal and external scientists [3]. The Dst tracked in this catalog is a record of geomagnetic activity from 1957 through 2007, which holds data from five solar cycles. Solar cycles, approximately 11-year-periodic changes in the Sun's activity measured by sunspots on the solar surface, are relevant to the context of this work due to the independent and identical distribution (I.I.D.) assumption often required for the following type of analysis. One point of interest to consider when analyzing the Dst measurements provided by this catalog is that raw data is not presented but rather a post-processed version of the values. Both smoothing functions and frequency domain filtering are applied to the time-series to account for confounding phenomena such as the Moon's orbit, the Earth's rotation, and mutual coupling of the same. The following work is done using absolute values of Dst, further referred to as |Dst|, for convenience.

Methods

The initial step in extreme value analysis is sampling the extreme values from the data. There are two main approaches used to select maxima from a time-series: the block maxima method extracts the largest value in a predefined time window, and the peak over threshold method accepts all values greater than a predefined threshold.

Block Maxima Method

The block maxima method for sampling extreme values typically produces a set of data that can be characterized by the Generalized Extreme Value distribution (GEV). The GEV is a combination of three common extreme value distributions, the Gumbel, Frechet, and Weibull distributions. This formulation allows the data to automatically choose the distribution that best describes it through the values found for the GEV's parameters during parameter search. The probability density function (PDF) is of the form in Equation 1.

$$f(\zeta,\ \mu,\ \sigma) = exp\left(-\ [1\ +\ \zeta(\frac{x-\mu}{\sigma})]^{-\frac{1}{\zeta}}\right) \qquad (1)$$

Where the shape parameter, ζ, provides the tail behavior, the location, μ, is the |Dst| average, and the scale, σ, is a measure of variability in the data. Multiple time windows are explored in this work to define the block within which maxima are selected. This parameter is a heuristic that is a point of weakness for the approach. A time window too large will deplete the number of samples and increase uncertainty in the model while a window too large will accept too many data points and increase bias in the model. The distribution fitting was accomplished using Maximum Likelihood Estimation (MLE) where the joint probability of the data is maximized with respect to the parameters using a numerical optimizer. The SciPy Python library was deployed to compute the parameters for the GEV [4].

Peak Over Threshold Method

The peak over threshold approach typically corresponds to the Generalized Pareto Distribution (GPD). The GPD is another distribution that is commonly used to model extreme values where it also allows the data to influence the specific type of distribution that is found through parameter fitting procedures. Depending on the parameter values found, the distribution may become an exponential or Pareto distribution, which can both be used to model exceedances. The PDF is of the form in Equation 2.

$$f(\zeta,\ \mu,\ \sigma) = 1\ -\ [1\ +\ \zeta(\frac{x-\mu}{\sigma})]^{-\frac{1}{\zeta}} \qquad (2)$$

Where each parameter represents the same information as in the GEV distribution. MLE using the previously mentioned Python library is again used to compute the distribution parameters. In this case the threshold is the heuristic that introduces weakness into the approach. If a threshold too high is selected, too few samples will be included and this will again introduce uncertainty into the model while a threshold too low will accept too many data points and undermine the basis of distributions that model extreme values. Like block maxima window selection, threshold selection is more of an art than a science, however, there are techniques to semi-quantitatively evaluate the optimal threshold before evaluating the resulting probability model. The mean excess function, described in Equation 3, is the expectation on the data given the data exceeds a specified value, and by observing its behavior across all thresholds, it is possible to pick a threshold that will potentially avoid the pitfalls associated with a threshold heuristic.

$$e(x) = E(X > x) = \frac{\int_x^{\infty}(1-F(u))du}{1-F(x)} \quad (4)$$

Where x is the threshold, $u = X - x$, and F is the cumulative distribution function (CDF). In practice the mean excess function can be estimated using a sample of data, and this estimation is given by Equation 5.

$$\hat{e}_n(x) = \frac{\sum_i^n (x_i - x) I_{x_i > t}}{\sum_i^n I_{x_i > t}} \quad (5)$$

Where n is the number of data points in the sample. By plotting the value of $\hat{e}_n(x)$ and a confidence interval for $\hat{e}_n(x)$ over all thresholds, a range of values can be observed for which the mean excess function is relatively constant. A threshold in this range will likely generate a more reliable GPD model, and picking a value at the beginning of this range will help to decrease uncertainty because it will allow for more data points in the extreme value sample. A secondary effort that may increase confidence in the threshold selection is to evaluate parameter stability over all thresholds. At each threshold a GPD model may be fit to the resulting data. By plotting the shape and scale parameters along with their confidence intervals, parameter stability and sampling error can be evaluated for the range in which stability was found for the mean excess function.

Once parameter sets have been optimally computed for the respective sampling methods, storms of large magnitude can be forecasted for specified time periods. For a given time period T, the return level is defined as a |Dst| value that is expected to be reached or exceeded on average every T years with probability given in Equation (6).

$$P_{return} = \frac{1}{T} \quad (6)$$

The return level at the specified time can then be solved for by inverting the CDF of the distribution at P_{return}.

Results

Block Maxima Method

The approach taken in this work is a naive approach in which solar cycles are ignored so that all data from the set may be used in the analysis. This approach is favorable with respect to uncertainty in the resulting estimates as it allows us to use all available data. However, it disregards an important assumption for building the distributions, which is that all data are independent.

Two time windows are investigated for sampling maxima: one month and one year. Figure 1 displays the empirical and parametric PDFs for both time windows. Visual inspection of the plots confirms that the GEV is a reasonable distribution to fit the data. Histogram binning is accomplished using the Sturges and Freedman-Diaconis method as implemented in the Matplotlib Python library, which seeks to minimize the integral of the squared difference between the histogram and the theoretical density [5]. Manual binning may result in a biased evaluation of model selection since it is possible to visually search for a bin width that would look the most similar to the continuous density function.

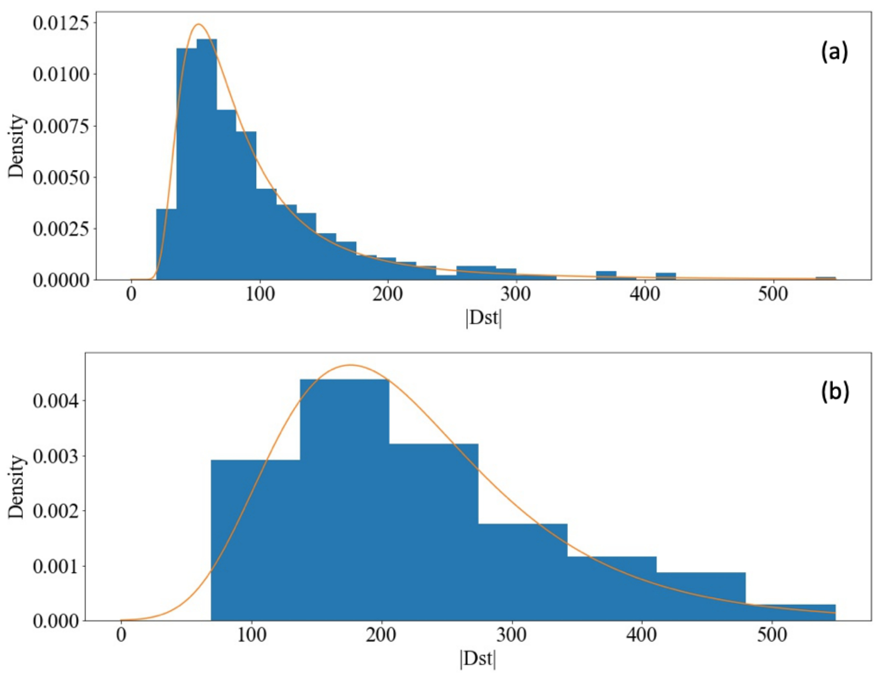

Figure 1. Generalized Extreme Value probability density function and empirical distribution (histogram) fit to extreme |Dst| values sampled using the block maxima method with time windows of (a) one month and (b) one year.

In addition to the distribution plots, we can observe the probability plots for the two time windows as well. The probability plots show the sample data versus the quantiles of the theoretical distributions. Figure 2 displays these plots. While the model fit with the one-month time window has a Pearson correlation of 0.938, it can be seen that the sample and theoretical values do not follow an exactly linear relationship. Importantly, it is also unable to accurately estimate extreme |Dst| values at the higher end of the scale. The model fit with a one-year time window has a Pearson correlation of 0.994 and visually seems to estimate extreme |Dst| values reasonably well across values, however, a slight oscillatory trend around the least squares curve is evident, which may suggest some systematic bias in the model.

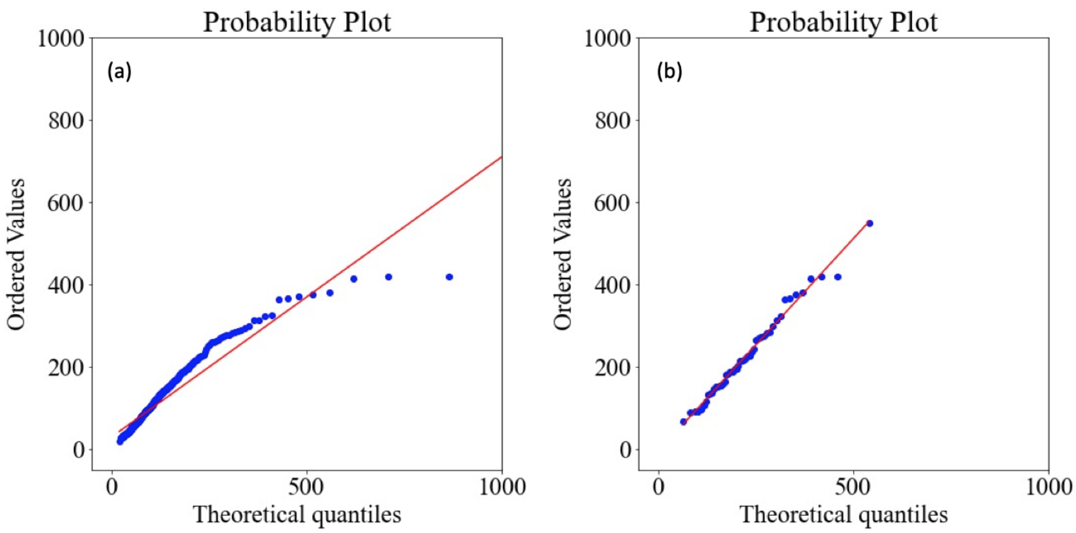

Figure 2. Probability plots with least squares regression lines showing the agreement between the Generalized Extreme Value distribution fit to the data and the sample values themselves where the fit distributions are built with extreme values sampled from the block maxima method using time windows of (a) one month and (b) one year.

Given the better fit of the model based on annual block maxima extreme values, return levels are computed for only this model. Return levels are shown in Figure 3. The parameters found with MLE for this model are: $\zeta =- 0.03$, $\mu = 178.63$, and $\sigma = 79.29$ where the negative shape parameter indicates that there is an upper bound on the |Dst|, and this can be confirmed in the return level plot by observing the plateau behavior of the curve [6].

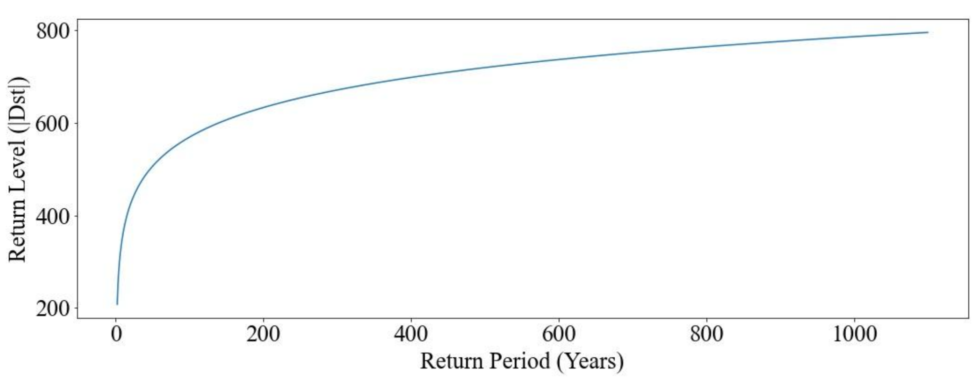

Figure 3. Return level calculated between 0 and 1100 years using the Generalized Extreme Value distribution generated by annual block maxima extreme value sampling.

Table 1 provides the estimates of |Dst| return level after 10, 50, and 100 solar cycles. Code was written to compute non-parametric bootstrapped 95% confidence intervals, however, this requires a large number of optimization runs, so this information was not generated due to the time intensive computation required.

	110 years	550 years	1100 years
|Dst| Return Level	577.88	728.25	795.14

Table 1. |Dst| return levels estimated after 10, 50, and 100 solar cycles using the Generalized Extreme Value distribution.

Peak Over Threshold Method

The naive approach with respect to independent data is again taken with the peak over threshold method. As described in the Methods section, a mean excess plot is generated to identify a suitable threshold for sampling extreme values from the total time-series. Figure 4 visualizes the mean excess function across all possible thresholds.

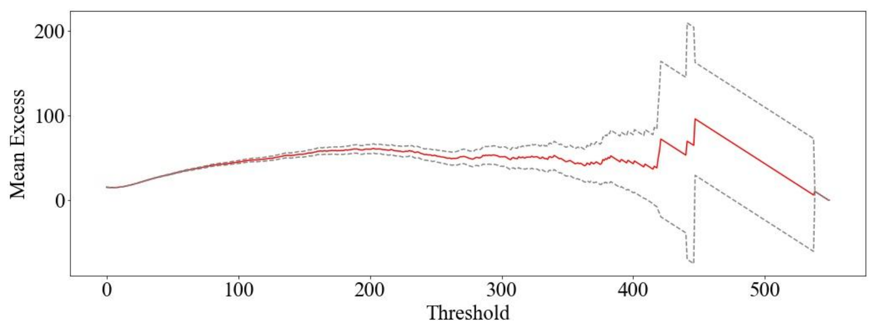

Figure 4. Mean excess function plotted across all possible thresholds where the gray dashed line represents a 95% confidence interval.

As expected the confidence interval widens with decreasing data (i.e., higher thresholds). The idea is to select a threshold where the mean excess function is constant and where the confidence interval has a tighter bound. Based on these criteria, a threshold of 260 is chosen. To confirm the validity of this threshold, parameter stability plots are also generated. Figure 5 displays the GPD shape and scale parameters across thresholds.

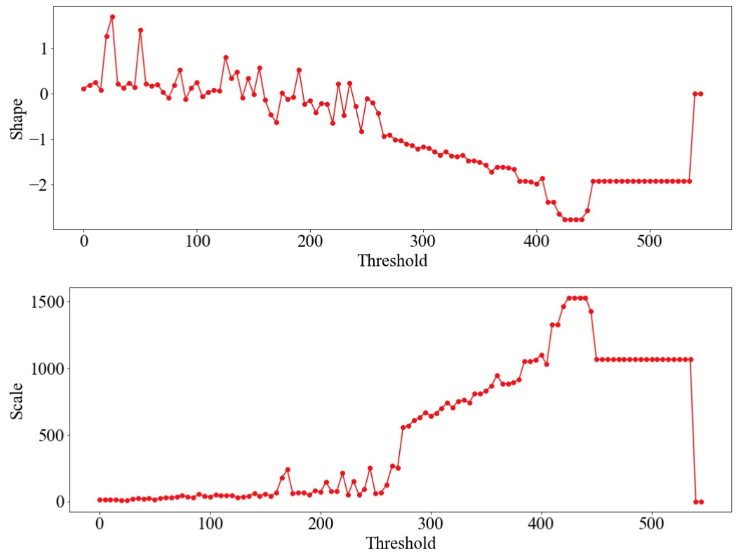

Figure 5. Generalized Pareto Distribution parameter stability plots generated over |Dst| thresholds computed in intervals of five.

Distributions are fit in intervals of five to reduce computation time. Code was also written to generate non-parametric bootstrapped 95% confidence intervals, however, this would require thousands of distributions to be fit at each threshold over hundreds of thresholds, which is too computationally expensive. A location parameter plot is not required because this value will simply increase linearly as data points are deleted from the sample in increasing value. While the shape and scale are not constant across the same range of thresholds as the mean excess function, they are not noticeably unstable, so we will continue with the original threshold of 260.

Given this threshold the parameters found using MLE are: $\zeta = -0.43$, $\mu = 260.20$, and $\sigma = 124.51$. The PDF and empirical distribution plot can be created with the parameters found similar to those in the block maxima method section (Figure 6).

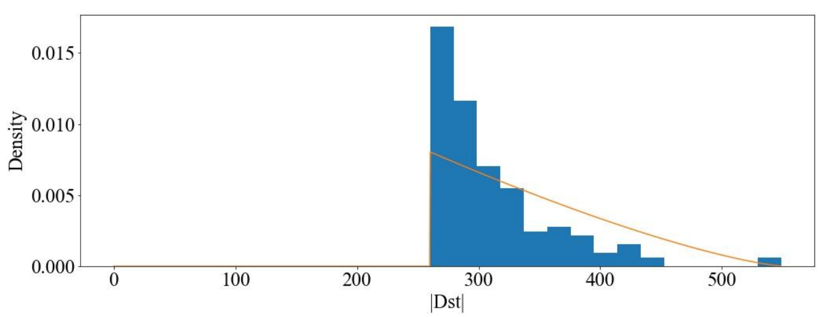

Figure 6. Generalized Pareto Distribution probability density function and empirical distribution (histogram) fit to extreme |Dst| values sampled using the peak over threshold method.

The histogram seems to suggest that a GPD model is a reasonable choice, however, the distribution fit from the data underestimates the probability of lower extreme values. To further evaluate the model fit, the probability plot is observed in Figure 7.

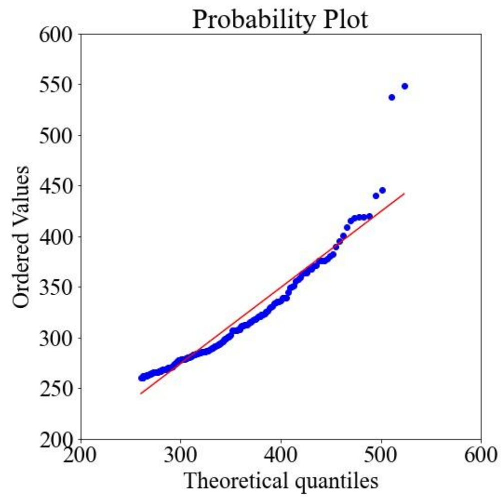

Figure 7. Probability plot with least squares regression line showing the agreement between the Generalized Pareto Distribution fit to the data and the sample values themselves where the fit distributions are built with extreme values sampled from the peak over threshold method.

The theoretical and sample values have a Pearson correlation of 0.956, but there is a slightly non-linear trend between the data.

Return levels for this sampling method are computed and shown in Figure 8. With respect to the return level curve plotted for the block maxima method, this return level curve seems to plateau more quickly, which suggests that we are more likely to experience extreme geomagnetic storms sooner than the block maxima method would estimate.

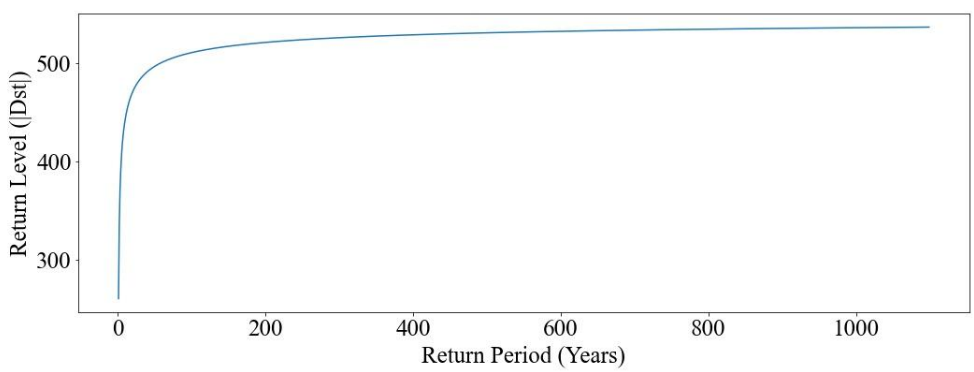

Figure 8. Return level calculated between 0 and 1100 years using the Generalized Pareto Distribution generated by peak over threshold sampling.

Table 2 provides the same information as in Table 1 but for the peak over threshold method.

	110 years	550 years	1100 years
|Dst| Return Level	511.97	531.28	536.27

Table 2. |Dst| return levels estimated after 10, 50, and 100 solar cycles using the Generalized Pareto Distribution.

Discussion

In this work I have explored two methods for sampling and modeling extreme values from a time-series to forecast the time-dependent severity of future geomagnetic storms. From visual observation of distribution plots, it seems that the block maxima method was able to produce a probability model that better fit the sample data. The block maxima method also estimated higher return levels than the peak over threshold method for the specified time periods. Both methods forecast geomagnetic storms of greater magnitude than those mentioned in the Introduction section in approximately the next 100 years.

As discussed previously the complete time-series used in this analysis does not contain all independent data points, so future work could explore the forecasts given data split by solar cycle. However, this brings to the forefront a weakness of the block maxima method, which is that it generates insufficient extreme value sample sizes given a time-series with a shorter timespan. A solar cycle is approximately 11 years, so block maxima sampled annually would provide only 11 data points with which to fit the GEV, increasing uncertainty in the resulting distribution. The peak over threshold method may be more reliable under these circumstances as it would likely accept more extreme values in the sample.

Code available here: https://github.com/isaacfinberg/Probability-and-Stochastic-Processes-2

References

1. Chen, S., Chai, L., Xu, K., Wei, Y., Rong, Z., Wan, W. (2019). Estimation of the occurrence probability of extreme geomagnetic storms by applying extreme value theory to aa index. Journal of Geophysical Research: Space Physics, 124, https://doi.org/10.1029/2019JA026947.
2. "The Disturbance Storm Time Index." *NOAA National Centers for Environmental Information*, NOAA, www.ngdc.noaa.gov/stp/geomag/dst.html.
3. U.S. Geological Survey (2009). USGS Geomagnetism Program: Definitive Disturbance Storm-Time (Dst) Data: U.S. Geological Survey data release, https://doi.org/10.5066/P9V6WO7G.
4. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272. https://doi.org/10.1038/s41592-020-0772-5.
5. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007. https://doi.org/10.1109/MCSE.2007.55.
6. Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer. https://doi.org/10.1007/978-1-4471-3675-0.