Extreme Value Theory and Regression Trees

Morgan Huang

Project

- Analyse ERA5 reanalysis surface wind speeds at 1000hPa in Toronto Pearson using Extreme Value Theory
  - Data from 1979-2020
- Return levels (future estimates) will be used to validate whether timber wood can withstand wind extremes

Structure:

- General Theory of Extremes
- Regression Tree implementation
- Results

Modeling Non-stationary Processes of Extremes using Regression Trees

Morgan Huang [1] Paul Kushner [2] Karen Smith [3]

[1]Department of Computer and Mathematical Sciences, University of Toronto Scarborough [2]Department of Physics, University of Toronto [3]Department of Physical and Environmental Sciences, University of Toronto Scarborough

Abstract

We focus on modeling extreme-valued wind speeds which can help us provide statistical estimates on future wind speeds. Typically, extremal data can be easily modeled using extreme value theory (EVT), but usually assume stationary data. Modeling non-stationary processes, which are common in environmental applications, however, is a non-trivial task. In this study, we explore extreme wind events in Toronto between 1979-2020 using ERA5 reanalysis data, and aim to use regression trees to form quasi-stationary clusters, for which we use a point process approach with regression parameters using the clusters as binary variables to provide statistical conclusions on its return values.

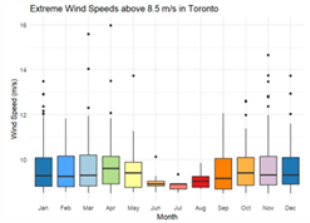


Figure 1: Boxplots of Extreme Wind Speeds showcasing seasonal cycle

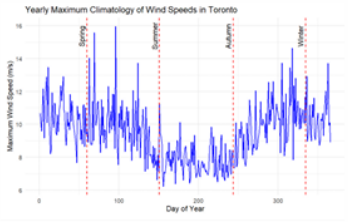


Figure 2: Daily Maximum Climatology of Wind Speeds between 1979-2020

Atmospheric Covariates

We would like to use covariates in our extreme value analysis of extreme winds not only to form stationary slices of our non-stationary data, but to also better understand what causes extreme wind events. To determine these covariates, we want to understand physical relationships between atmospheric variables and wind speed, and potential causes of extreme wind events.

Thermal Wind and Atmospheric Stability

In theory, wind is created by horizontal pressure gradients, where areas of high pressure exert some level of force towards areas of low pressure. From Figure 3, colder temperatures lead to a stronger decrease in pressure with respect to altitude than hotter temperatures, resulting in pressure gradients at different vertical altitudes. We calculate thermal wind, which uses temperature gradients to calculate the rate of change in horizontal wind speeds (U_g, V_g) with respect to altitude z:

$$\frac{\partial U_g}{\partial z} = -\frac{|g|}{T_v \cdot f_c}\frac{\partial T_v}{\partial y} \qquad \frac{\partial V_g}{\partial z} = \frac{|g|}{T_v \cdot f_c}\frac{\partial T_v}{\partial x}$$

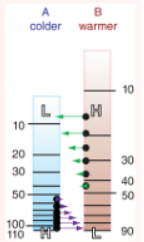


Figure 3: Difference in pressure is due to the hypsometric equation [7]

In addition to temperature gradients, an unstable atmosphere is heavily correlated with extreme wind events. Under unstable conditions, air parcels continue to rise to the top of the troposphere, which forces the jet stream winds to mix with surface winds. We can simply calculate the atmospheric stability by calculating the potential temperature gradient with respect to altitude.

Using covariates of atmospheric stability calculated in two layers (1000hPa-850hPA, 850hPa-500hPa), thermal wind gradients in the x and y direction, low level winds (winds at 850hPa), and the jet stream (winds at 500hPa), we fit a regression tree to find suitable discrete clusters. This makes our return level calculation easier and provides better information on the behavior of extreme winds without using interaction effects.

Methodology

We propose a point process approach in modeling extreme wind events. This is to overcome certain caveats that are present in other methods such as the block maxima approach, which is often wasteful of extremal data, and the peaks-over-threshold (POT) approach, where modeling non-stationarity often results in violating the threshold stability property (Eastoe and Tawn, 2009).

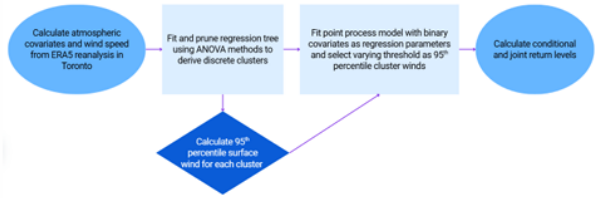


Figure 4: Flow Chart indicating the steps in fitting extreme value winds to a GEV distribution using the point process approach. We use ANOVA to fit our regression tree, but better theoretical basis can be found by setting the objective function as the negative log-likelihood (Farkas et. al 2024).

Results

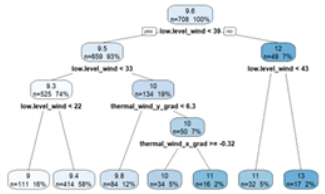


Figure 5: Regression Tree

After fitting and pruning the regression tree using ANOVA methods on extreme wind events (wind speeds of over 8.5 m/s), we obtain 7 different clusters, as shown in Figure 5. Thresholds for each cluster is heuristically chosen as the 95th percentile surface wind (Figure 6). Figure 7 shows the varying threshold with surface winds. Points above the threshold are considered extreme given the cluster, and are fitted into the point process model.

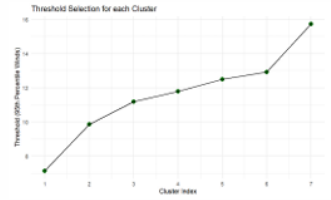


Figure 6: Threshold Selection (95th percentile) for each cluster

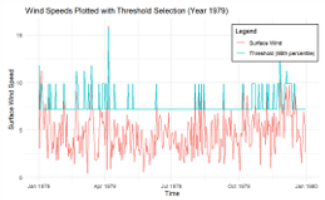


Figure 7: Plot of Varying Threshold and Wind Speed in the year 1979

We can perform model diagnostics on our point process model using a QQ-plot as shown on the right. Since our parameters vary with time, standardization is required. In this case, we convert the GEV parameters to parameters of a Generalized Pareto distribution, then standardize to a standard exponential distribution and plot its theoretical quantiles.

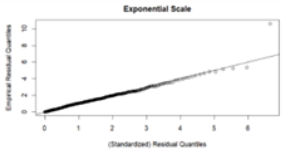


Figure 8: QQ-plot matches almost perfectly with the exception of one outlier

Return Levels

Working with a non-stationary process, we can immediately calculate conditional return levels $z_{p,k}$, where the probability that $z_{p,k}$ is exceeded for a given year is given by p, conditioned on the fact that all data points in a given year belong to cluster C_k:

$$\mathbb{P}(Z > z_{p,x} | Z \in C_k) = p$$

If we are interested in general return levels regardless of cluster partitions C_k, then we can calculate z_p by solving the following equation:

$$\mathbb{P}(Z > z_p) = \sum_{k=1}^{m} \mathbb{P}(Z > z_p | Z \in C_k)\mathbb{P}(Z \in C_k) = p$$

We thus calculate conditional return levels for each cluster k along with its bootstrap 95% confidence intervals on the left, and then we solve the general return level equation with 95% bootstrap confidence plotted on the right:

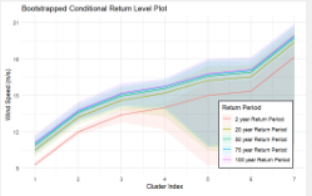


Figure 9: Conditional Return Levels for each cluster for return period of 2, 20, 50, 75, and 100 years

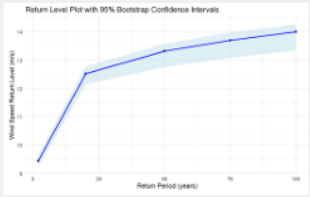


Figure 10: General Return Level plots

Assuming that the 40-year historical distribution of wind speeds do not change in the future, we expect that a wind speed of 14 m/s will be exceeded once every 100 years.

Discussion

Assuming that the historical distribution will be equal to future distributions of wind speeds is a strong assumption, especially with regards to climate change (Milly et al. 2008). Since extrapolation to future processes is not favorable (Cooley, 2013), environmental studies must focus on providing EVT analysis on not only historical, but future climate model projections. In this study, we were only able to provide information about historical extremes, since regular climate models (resolution of around 1 degree latitude and longitude) show no obvious trend in extreme wind speeds. However, (Morris et. al 2024) recently showed some increases in extreme wind speeds in Southern Ontario, by using a "variable-resolution" grid which increases resolution towards the area of study. This paper showed that extreme wind events are not caused by global effects of extratropical cyclones, but rather from locally reduced atmospheric static stability which cannot be detected by regular climate models.

Acknowledgments

This project is supported by the Data Sciences Institute, University of Toronto. The author thanks Paul Kushner and Karen Smith for their guidance on this project.

References

[1] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001. [2] Daniel Cooley. Return periods and return levels under climate change. *Extremes in a Changing Climate*, pages 97–114, 2013. [3] Jonathan A. Tawn Emma F. Eastoe. Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 58(1):25–45, 2009. [4] Michael Morris et al. Resolution dependence of extreme wind speed projections in the great lakes region. *Journal of Climate*, 37:3153–3171, 2024. [5] Paul CD Milly et al. Stationarity is dead: Whither water management? *Science*, 319:573–574, 2008. [6] Sebastien Farkas et al. Generalized pareto regression trees for extreme event analysis. *Journal of Extremes*, 27:437–477, 2024. [7] Roland Stull. *Practical Meteorology: An Algebra-based Survey of Atmospheric Science*. Dept. of Earth, Ocean Atmospheric Sciences, University of British Columbia, 2017.

https://www.linkedin.com/in/morganhuang237/ SUDS Showcase 2025, 10th floor, 700 University Avenue, Toronto morgan.huang@mail.utoronto.ca

Block Maxima Approach

Suppose $X_1, X_2, \ldots$ is sequence of independent and identically distributed random variables. Let $M_n = \max(X_1, \ldots, X_n)$, thus:

$$\Pr\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z)$$

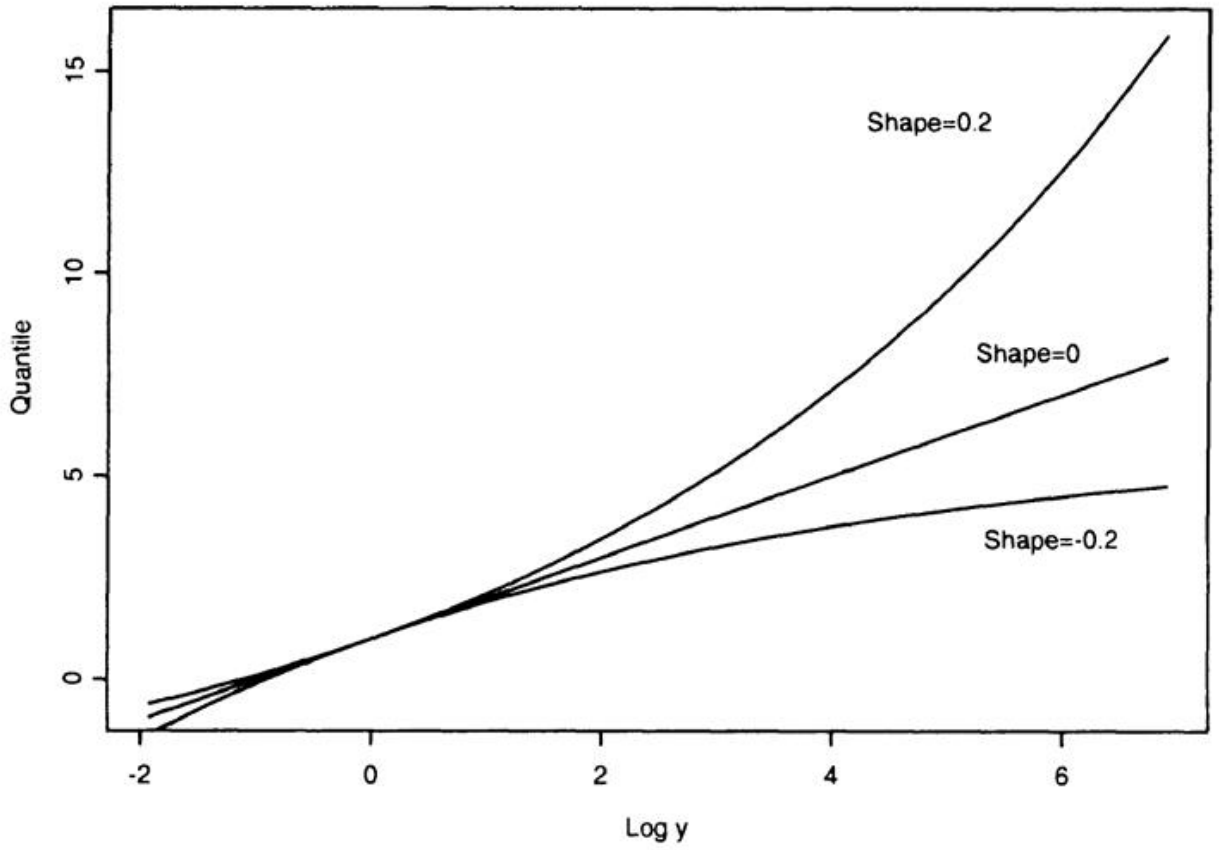
Where $\{a_n \geq 0\}$ and $\{b_n\}$ are sequence of constants, and $G(z)$ is a member of the GEV family, parametrized by μ, σ, ξ, the location, scale, and shape parameters, respectively.

$$M_n \sim GEV(\mu, \sigma, \xi)$$

- Fit annual maximas into model

Block Maxima Approach

- ξ, the shape parameter, controls the tail behavior
- If $\xi > 0$, then the distribution has 'heavy' tails
- If $\xi < 0$, then distribution has 'lighter' tails



Peaks over Threshold

Suppose $X_1, X_2, \ldots$ is sequence of independent and identically distributed random variables. For large enough threshold u:

$$\Pr(X - u | X > u) = 1 - (1 + \frac{\xi(X-u)}{\sigma})^{-1/\xi}$$

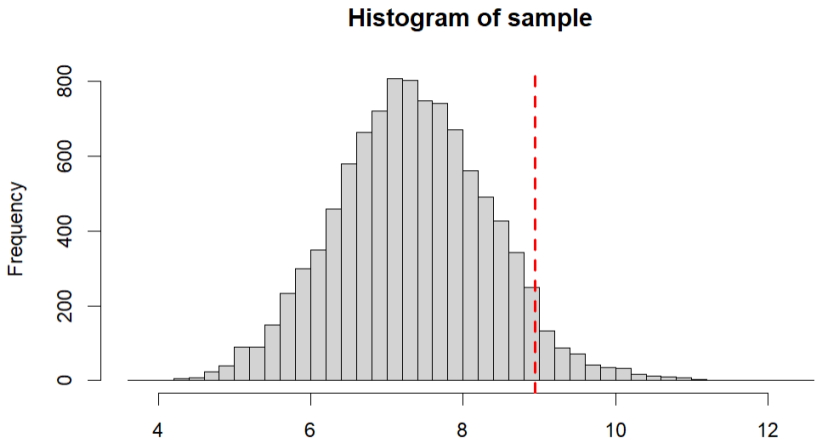
- Fit threshold exceedances above u

$$X - u | X > u \sim GP(\sigma, \xi)$$

- Big problem is choosing a threshold u, which is subject to bias-variance trade-off.
  - u is too large → model has a lot of variance
  - u is too small → model has a lot of bias

Bias-Variance Trade-off Simulation

- We perform a case-study simulation to showcase bias-variance trade-off:
- Sample from mixture distribution(Fréchet distribution on lower 95th tail, GP distribution on 5th upper tail), with 95th percentile as threshold selected
- Fréchet true parameters($\mu = 7, \sigma = 1, \xi = -0.3$)
- GP true parameters($\sigma = 0.6, \xi = 0$), this is what we want to estimate
- Fit to obtain sample parameters and repeat process 1000 times



Simulation Results

$$X - u | X > u \sim GP(\sigma = 0.6, \xi = 0)$$

	Mean σ	Standard Deviation σ	Mean ξ	Standard Deviation ξ
True Threshold	0.6	0.037	-0.038	0.045
High Threshold (True Threshold + 1)	0.56	0.085	-0.045	0.118
Low Threshold (True Threshold - 1)	0.7	0.016	-0.1	0.018

- High threshold corresponds to higher standard deviations, low threshold corresponds to high bias
- Simulation is not perfect, thus bias is still present in high threshold situation and even true threshold as well

Threshold Choice

Two ways of choosing a threshold. We formulate the relevant theory first.

Suppose $\mathrm{X} - u|X > u \sim GP(\sigma_u, \xi)$. Then, $E(X - u|X > u) = \frac{\sigma_u}{1-\xi}$.

Now, if $X - u$ is GP distributed, then $X - u_0$ should also be GP distributed, for $u_0 > u$(we just need to change the scale). Thus:

$$E(X - u_0|X > u_0) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0+\xi u}}{1-\xi}$$

Thus, $E(X - u|X > u)$ is a linear function of u.

Mean Residual Life Plot

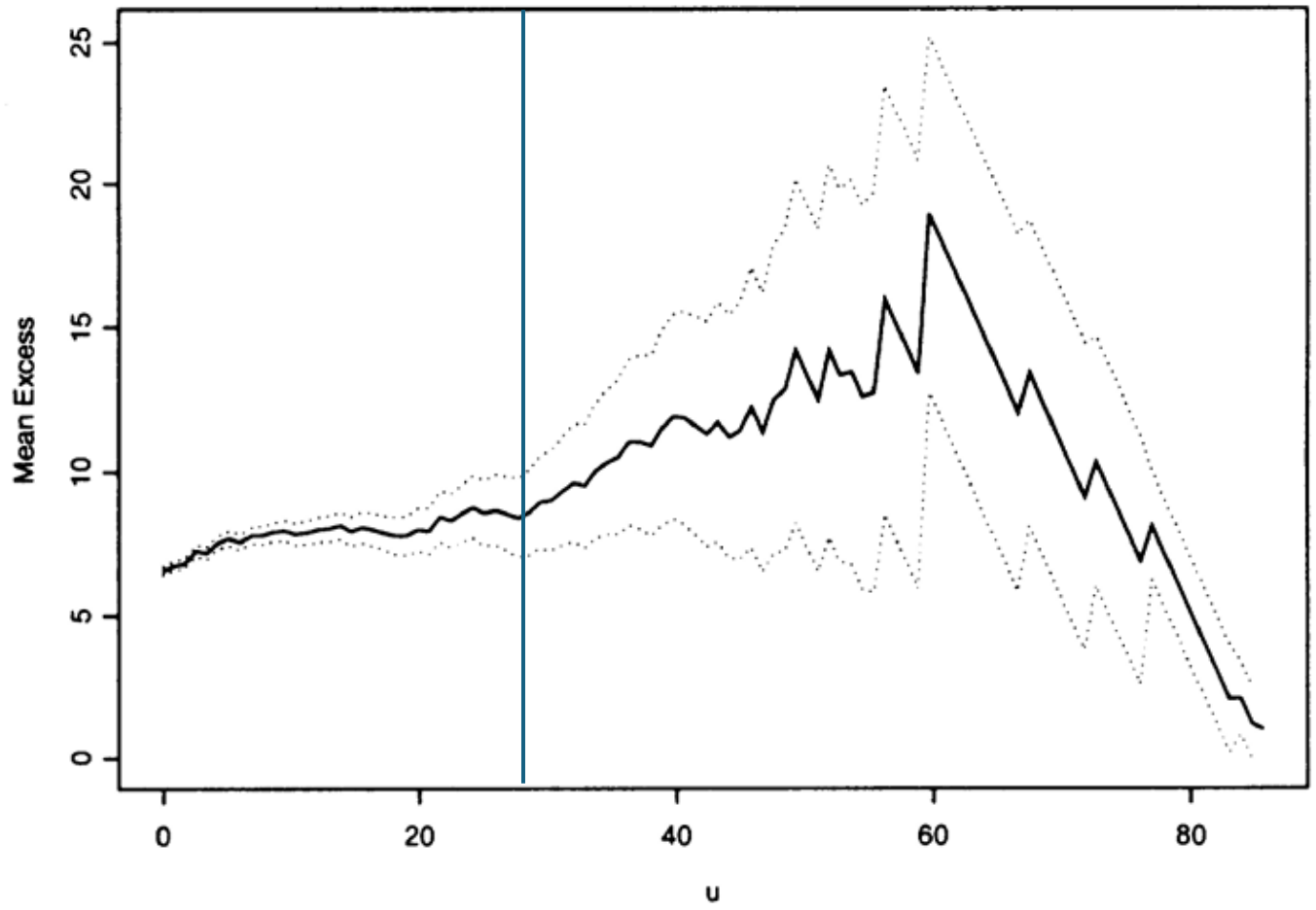
By the previous theory, if our model assumptions are correct, then $E(X - u|X > u)$ should change linearly with u. We can check this using empirical data by plotting the following:

$$\{(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_i - u)) \mid u < x\_max\}$$

Where n_u is number of threshold exceedances and threshold u must be below x_max.

Example

- We choose lowest u such that plot is approximately linear
- Take as much data as possible while keeping model assumptions true
- $u \approx 30$ is a good choice

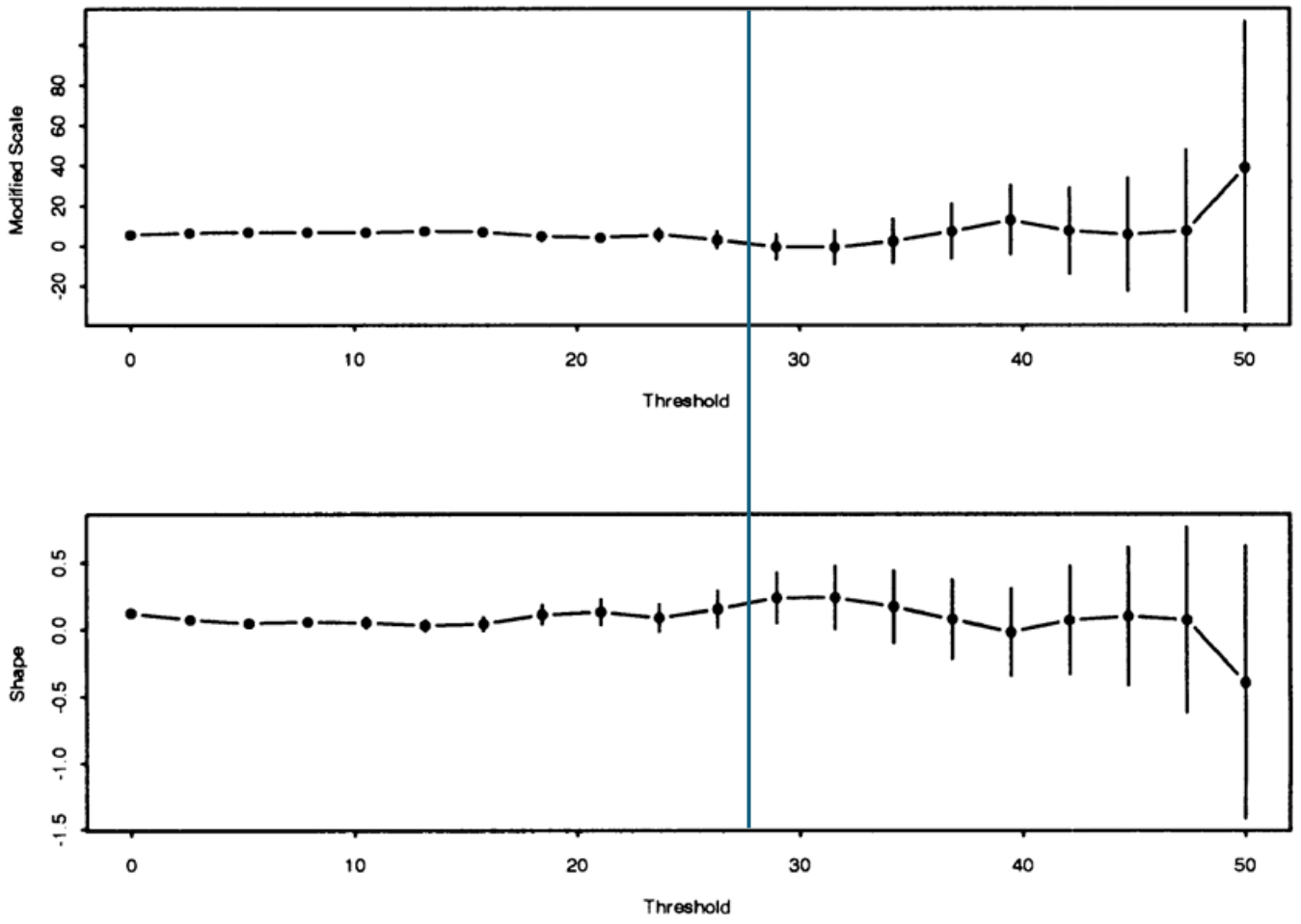


Threshold Stability

Another way to choose threshold is to find parameters that are stable across multiple thresholds. For threshold u and $u > u_0$:

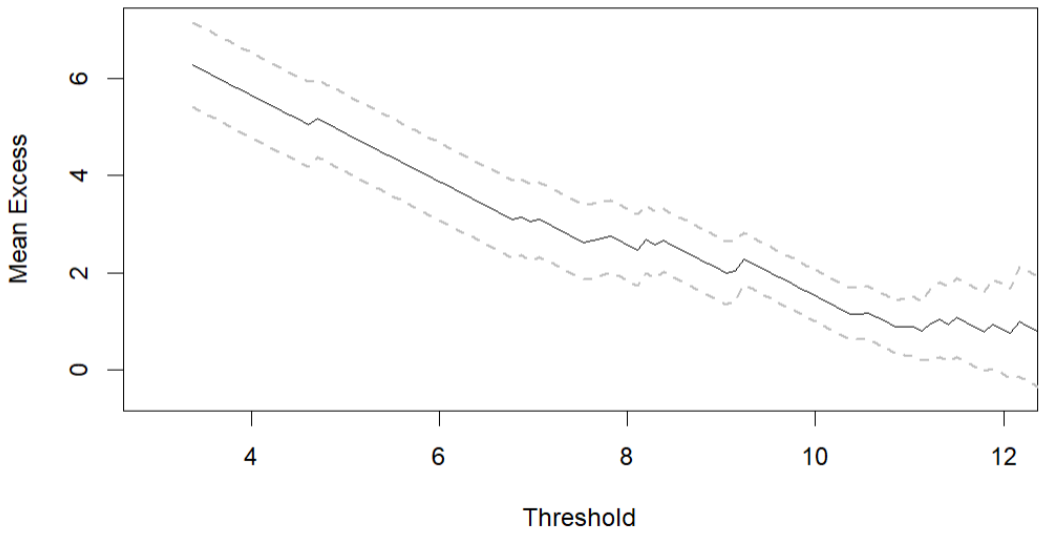
$$\sigma_u = \sigma_{u_0} + \xi(u - u_0)$$

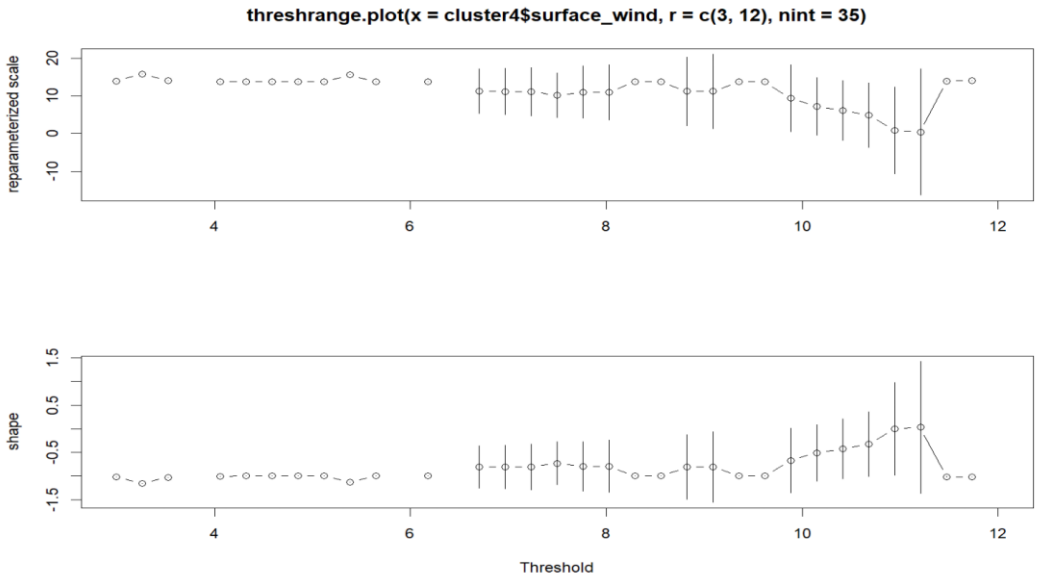
Reparametrize $\sigma^* = \sigma_u - \xi u$, then σ^* and ξ become constant under u, since $\sigma^* = \sigma_{u_0} + \xi u_0$. We plot σ^* and ξ with u and select lowest u for which the estimates are near constant



Problems with Threshold Selection

- Selecting threshold is very hard and subjective, as there are many candidates for thresholds and plots are often unclear
- Properties mentioned before are not exclusive
- For simplicity we heuristically choose 95th percentile for this project





Stationary and Non-Stationary Processes

- EVT Models assign the same probability distribution to every data point at every time step, thus it assumes stationarity

Def: A random process $X_1, X_2, \ldots$ is stationary if and only if given any set of integers $\{i_1, \ldots, i_k\}$ and $m \in \mathbb{Z}$ the joint distributions $\left(X_{i_1}, \ldots, X_{i_k}\right)$ and $\left(X_{i_{1+m}}, \ldots, X_{i_{k+m}}\right)$ are equal.

- Data points of a stationary process can be dependent of each other, but their distributions(marginal or joint) must be equal
- Many variables with seasonal cycles or trends are thus non-stationary

Non-Stationary Processes

Many ways of handling non-stationarity(Mentaschi et al, 2016):

- Fit regression parameters into extreme valued distribution

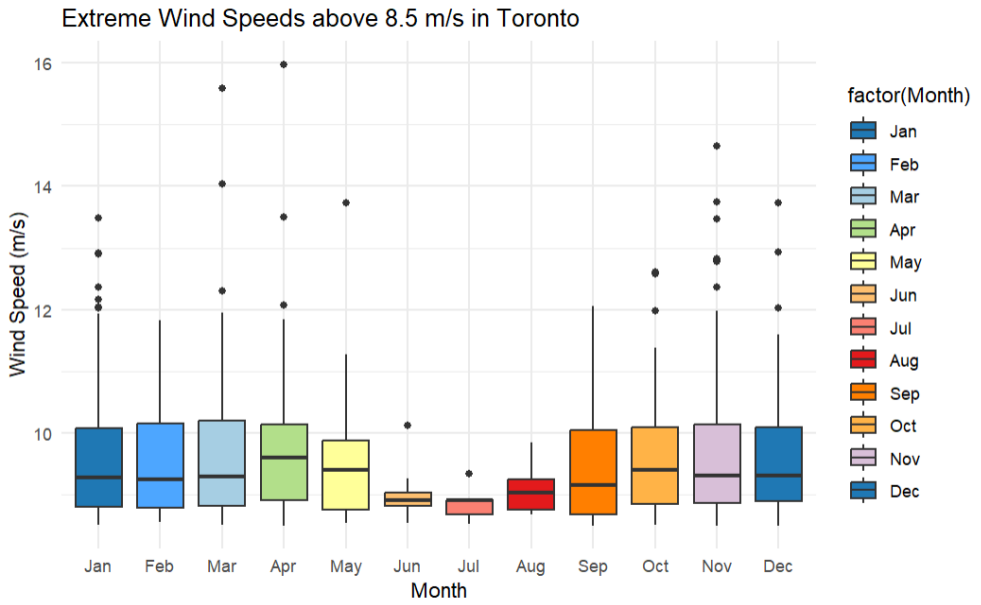
$$Y \sim GP\big(\sigma(t), \xi(t)\big)$$

where $\theta(t) = h(X^T\beta), \theta$ is any parameter, t is time

- Slice data into approximately stationary slices and fit separate models for each slice(ex. one model for each season)
  - Often wasteful of data
- Transform non-stationary process into stationary process through detrending and removing seasonality

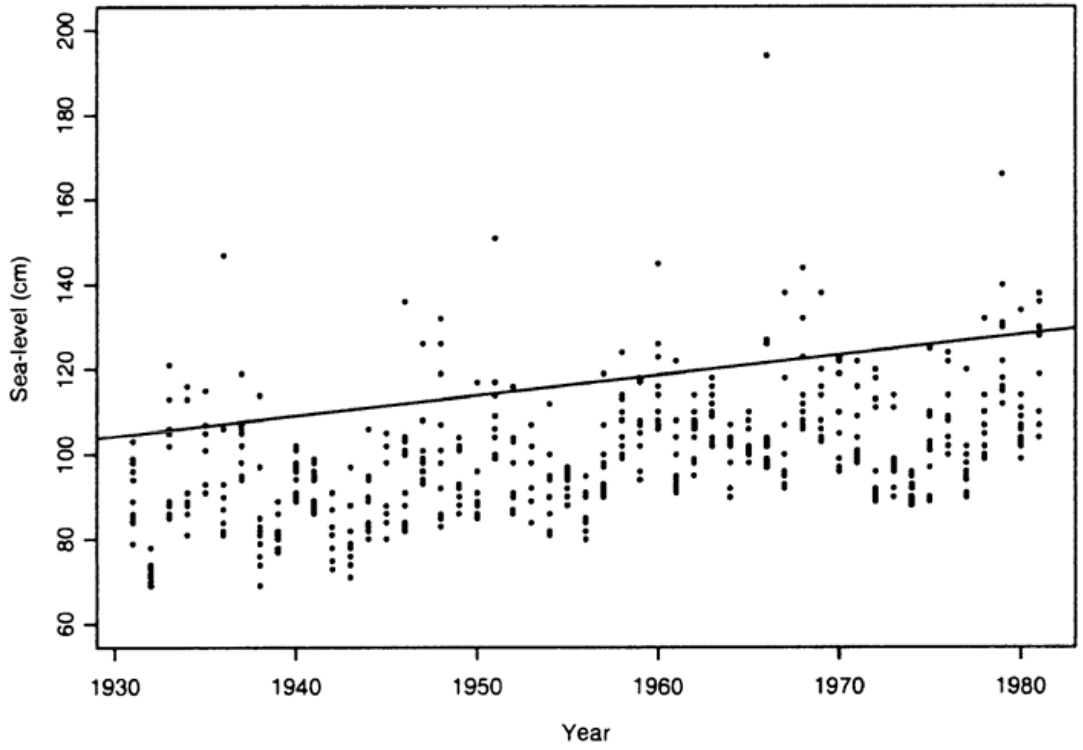
Methodology

- We will combine two strategies → Find approximately stationary slices, then create a categorical variable to represent each slice
  - Use categorical variable for regression parameters
- This will allow us to use all the data into the model fitting process



Threshold for Non-stationarity

- If we use threshold model, then there may be some problems:
  - Threshold may not apply across time
- Must have threshold function $u(t)$ as well
- Problem: When using threshold model with a varying threshold, problems in the model arises(Eastoe and Tawn, 2009)



In particular, the following does not hold for all covariates x:

$$\sigma_u(x) = \sigma_v(x) + (v - u)\xi(x)$$

Point Process Approach

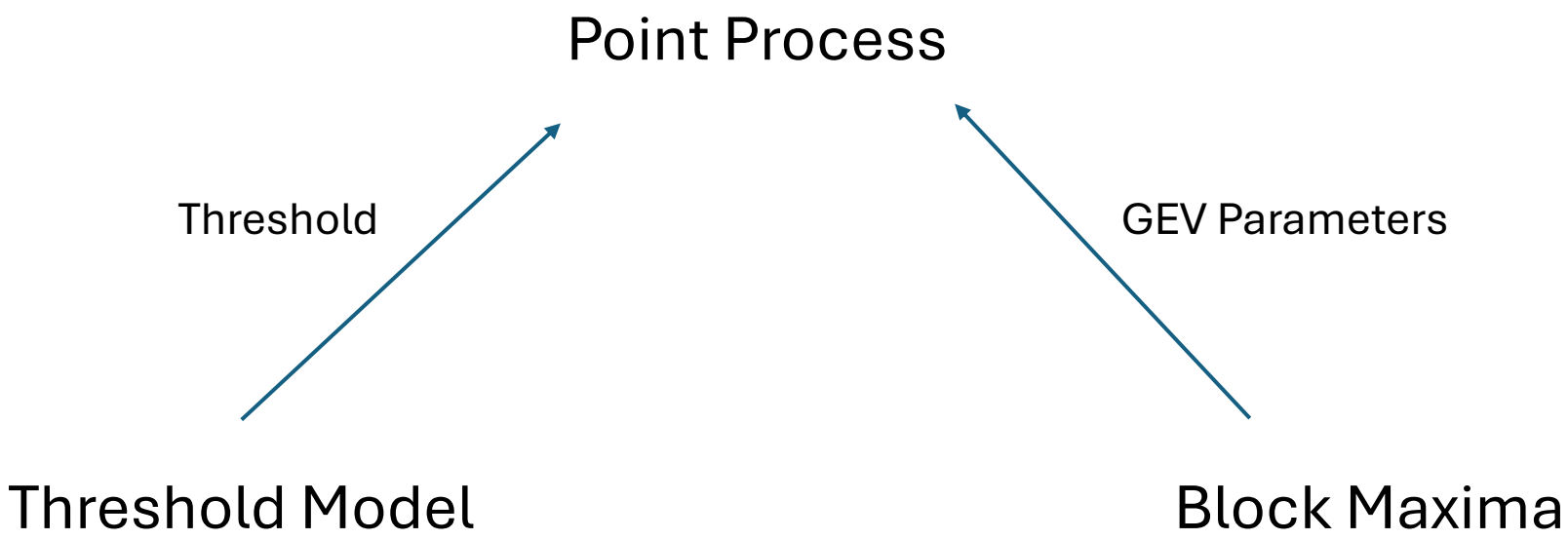
- To provide greater flexibility for varying threshold and to have its choice to have no effect on parameter estimates
- Extreme point occurrence in a set A(which can be defined using a threshold u) given by a Poisson process, with an intensity measure $\Lambda(A)$

$$N(A) \sim Poisson\big(\Lambda(A)\big)$$

Where $N(A)$ is number of points in A, and $\Lambda(A)$ is given in terms of GEV parameters:

$$\Lambda(A) = [1 + \xi\left(\frac{u-\mu}{\sigma}\right)]^{-1/\xi}$$

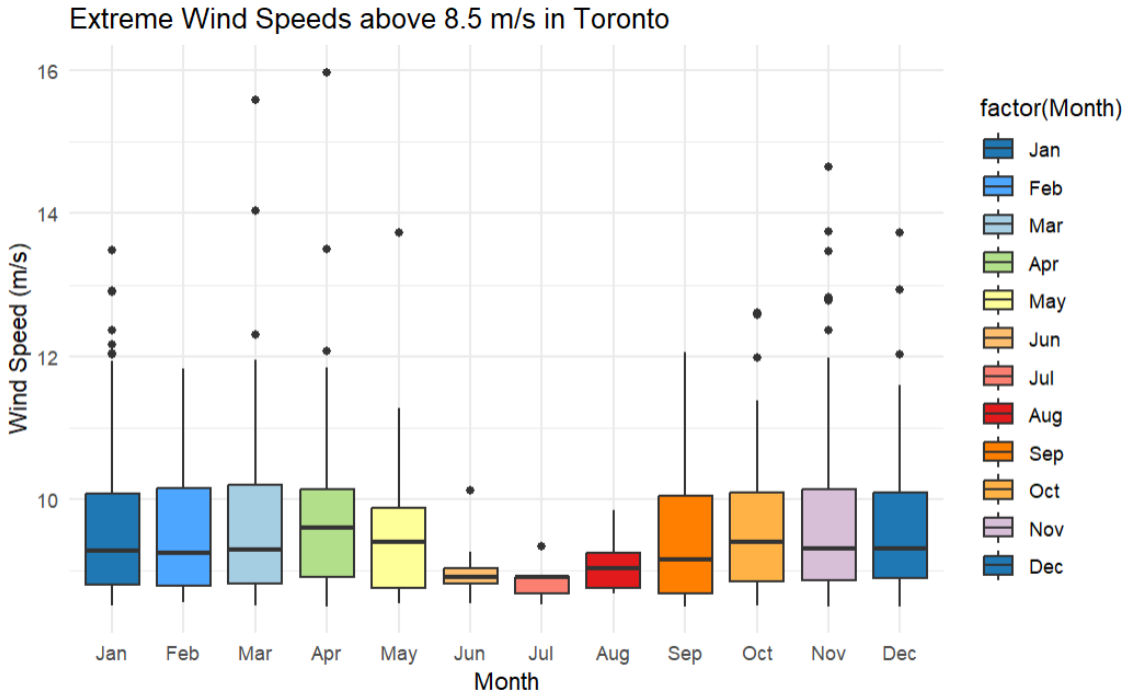
Point Process Approach



Threshold has no effect on the GEV parameters

Methodology

- We will combine two strategies → Find approximately stationary slices, then create a categorical variable to represent each slice
  - Use categorical variable for regression parameters and fit to a point process model
  - Define a varying threshold that depends on the category
    - 95th percentile for each category
- This will allow us to use all the data into the model fitting process



$$X = \begin{cases} 1, summer \\ 0, not\ summer \end{cases}$$

$$Y_u \sim GEV\big(\mu(X), \sigma(X), \xi(X)\big)$$

$$\begin{cases} \mu(X) = \mu_0 + \mu_1 X \\ \sigma(X) = \exp(\sigma_0 + \sigma_1 X) \\ \xi(X) = \xi_0 + \xi_1 X \end{cases}$$

Conditional and Joint Return Levels

- For stationary process, we solve for z_p:

$$\mathbb{P}(Z > z_p) = 1 - \exp\left\{-1\left[1 - \xi\left(\frac{z_p - \mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}\right\} = \mathrm{p}$$

- For non-stationary processes, we condition on a covariate X:

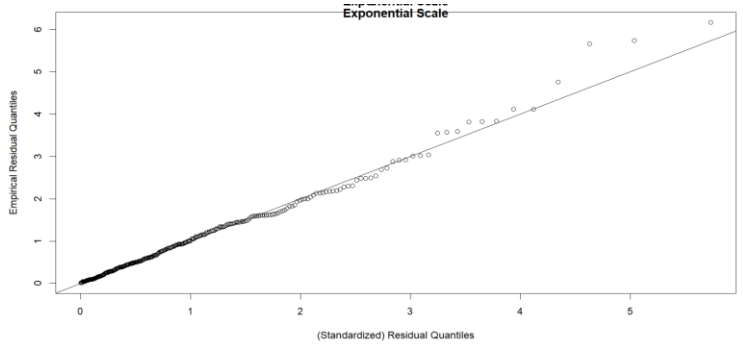
$$\mathbb{P}(Z > z_p) = \int_{x \in Dom(X)} \mathbb{P}(Z > z_p | X = x) f_X(x) dx$$

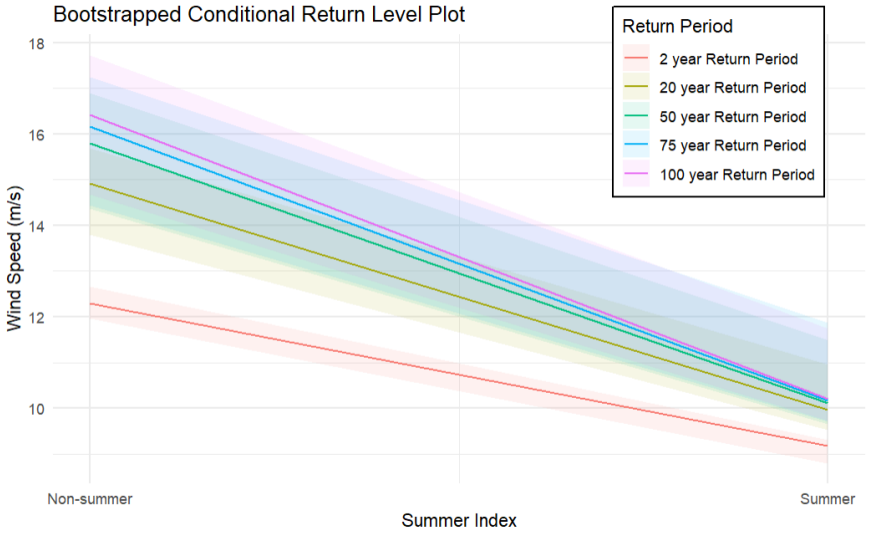
where $f_X(x)$ is the probability density of X

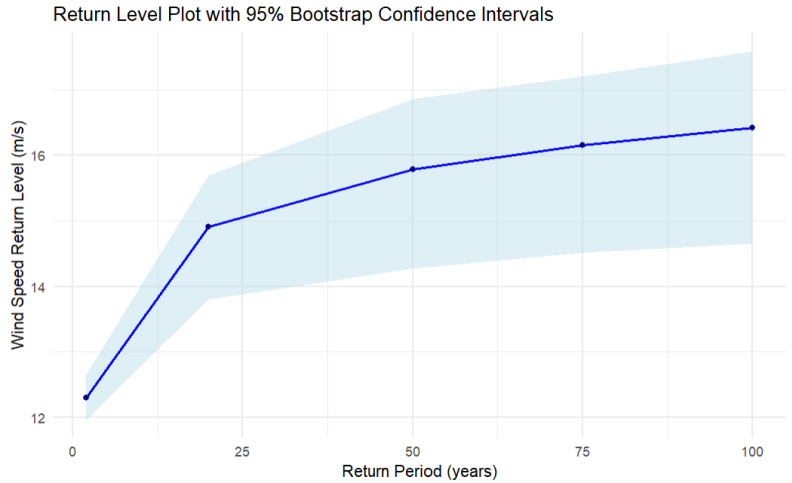
- We say that $\frac{1}{p}$ is the return period (in years), where we expect that z_p will be exceeded once every $\frac{1}{p}$ years

Toy Model with Summer Index

Parameter	Value	Std. Error
μ_0	11.906	0.157
μ_1	-2.909	0.209
σ_0	0.0719	0.088
σ_1	-0.701	0.127
ξ_0	-0.04	0.062
ξ_1	-0.31	0.114







Generalize to Multiple Covariates

Recall the return level calculation for one continuous covariate X:

$$\mathbb{P}(Z > z_p) = \int_{x \in Dom(X)} \mathbb{P}(Z > z_p | X = x) f_X(x) dx$$

- If we want to calculate for multiple covariates, then we would need double or triple integrals or more, which are much harder to calculate
- We revisit our toy model, where we separated data into discrete categories (binning)
- Solution: Find suitable clusters and use cluster as categorical variable

Cluster Approach

- Suppose we find clusters $C_1, \ldots, C_m$, for some $\mathrm{m} \in \mathbb{Z}$
  - Then, we define $\mathrm{m} - 1$ binary variables corresponding to each cluster except one which will be a dummy variable

$$X_k(t) = \begin{cases} 1, if\ Z_t \in C_k \\ 0, if\ Z_t \notin C_k \end{cases}$$

Then, GEV parameters can be defined as follows (ex. location μ):

$$\mu(\vec{X}) = \mu_0 + \mu_1 X_1 + \cdots + \mu_{m-1} X_{m-1}$$

Define varying threshold for each cluster X_k as its 95$^{\text{th}}$ percentile surface wind:

$$u(\vec{X}) = u_0 + (u_1 - u_0) X_1 + \cdots + (u_{m-1} - u_0) X_{m-1}$$

Where u_k is the threshold for kth cluster

Cluster Approach

More importantly, we have the following return level calculation which is a lot simpler:

$$\mathbb{P}(Z > z_p) = \sum_{k=1}^{m} \mathbb{P}(Z > z_p | Z \in C_k)\mathbb{P}(Z \in C_k) = p$$

For m clusters, where $\mathbb{P}(Z)$ is the proportion of clusters in the dataset.

Problem: How to define suitable clusters such that each cluster is approximately stationary?

- Moreover, we want to understand the uncertainty behind extreme winds and where it comes from, so clusters should be differentiable from each other

Regression Tree (Farkas et al. 2024)

- We propose a method from a paper called (Generalized Pareto Regression Trees for Extreme Event Analysis)
- The idea is to construct a regression tree which, for each iteration, will split the dataset into two branches, where the split minimizes some objective function:

$$m^*(x) = \text{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\phi\left(Z, m(X)\right)\right]$$

where ϕ is some loss function, $m \in \mathcal{M}$ is a class of target functions, Z is our variable of study, X is covariates.

- Two steps to regression tree → growing and pruning the tree
  - For simplicity we only focus on growing the tree since pruning requires cross-validation(will rely on heuristics)

Regression Tree Algorithm

- We apply the objective function to every leaf of the tree, starting from the root:

$$m^*(x) = \text{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\phi\left(Z, m(X)\right)\right]$$

- To grow tree, we split the root node into two using a rule $\mathcal{R}_l(X)$
  - Ex. For covariate X we can define rule $\mathcal{R}_l(X) = \begin{cases} 1, x < x_1 \\ 0, else \end{cases}$, for some x_1
  - Data points with covariate $x < x_1$ go to the left leaf of the root, and data points with $x \geq x_1$ go to the right
- The split criteria must minimize the total objective function:
  - Must satisfy:

$$\phi(Z, m(X)|Z \in C_0) - [\phi(Z, m(X)|Z \in C_1) + \phi(Z, m(X)|Z \in C_2)] < 0$$

where C_0 is the root, C_1, C_2 are the leaves

Regression Tree Objective Function

$$m^*(x) = \text{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\phi\left(Z, m(X)\right)\right]$$

- Farkas et al. proposes the negative log-likelihood of the GP distribution to be used as the objective function $\phi\left(Z, m(X)\right)$
- Thus, to find a suitable split, we do a grid search algorithm (iterate through all possible covariate values) and test each split
  - For each possible split created, we fit a separate EVT model for each leaf created (to find parameters $m = \theta$ through MLE methods)
  - Choose split that minimizes the negative log-likelihood
  - Very computationally expensive (took two days to train a regression tree)
- At the end of the growing and pruning process, the leaves of the tree represent the clusters

Problems

Many pros and cons to using negative log-likelihood function as objective function:

Pros:

- Strong theoretical basis
  - Far away points does not necessarily mean that it belongs to a different cluster, unlike standard ANOVA methods
  - Supports large shape parameters

Cons:

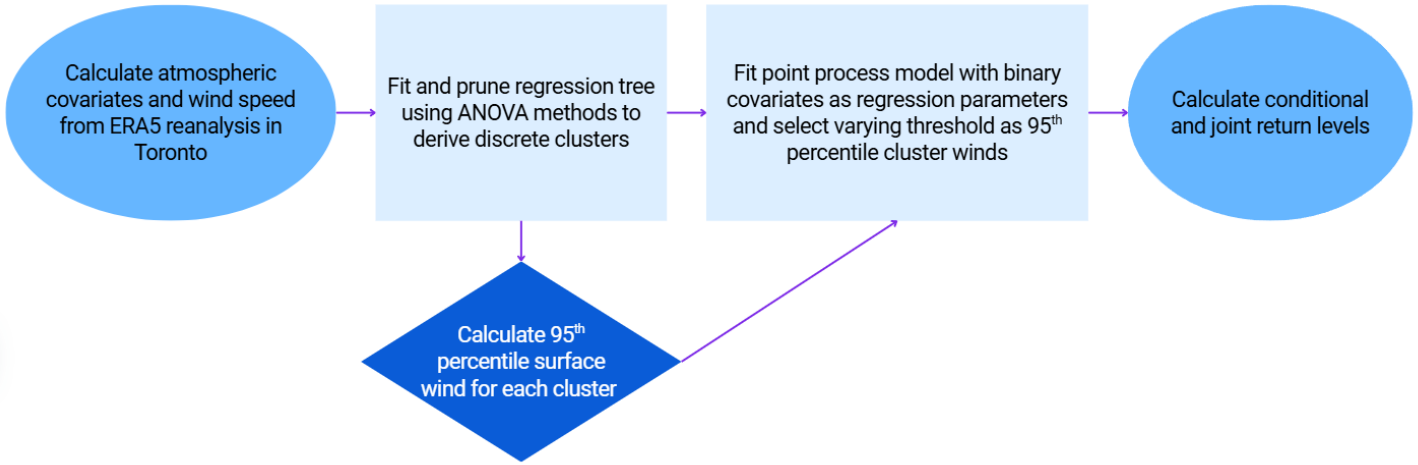
- Computationally expensive
- Must override package rpart source code
  - Negative log-likelihoods can often be negative, which is not supported
  - Farkas adds an artificial value that is dependent on the number of points in a leaf, but this is often arbitrary and hard to find suitable value

Problems

- To use ANOVA instead of negative log-likelihood
  - ANOVA calculates squared deviances from the cluster mean as its objective function
  - ANOVA prefers to minimize variance of each cluster, does not support large shape parameters

Overall Methodology

- Just like the summer index example, we do the same except now for multiple clusters
- We only fit data with wind speeds above 8.5 m/s to obtain the regression tree
  - Only want to characterize extreme winds not regular winds

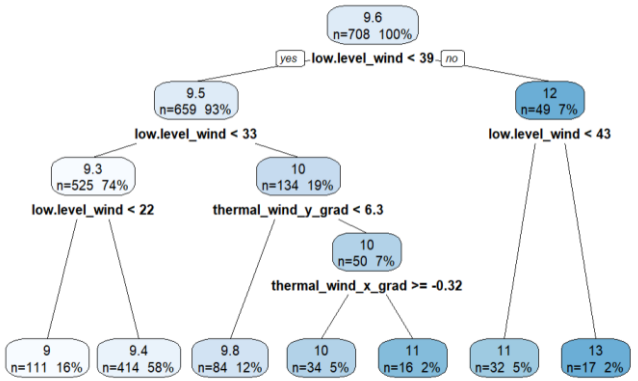


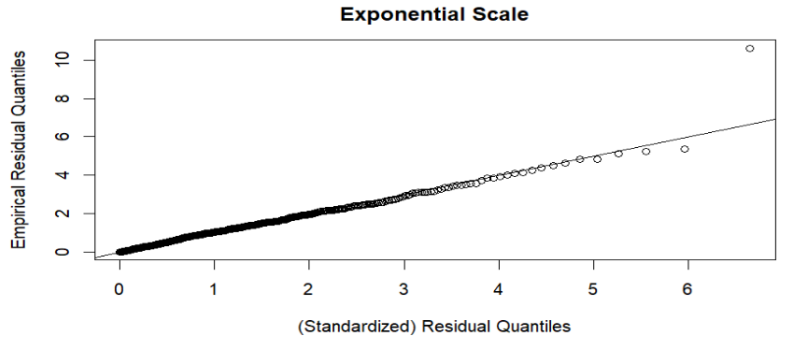
Atmospheric Covariates Used

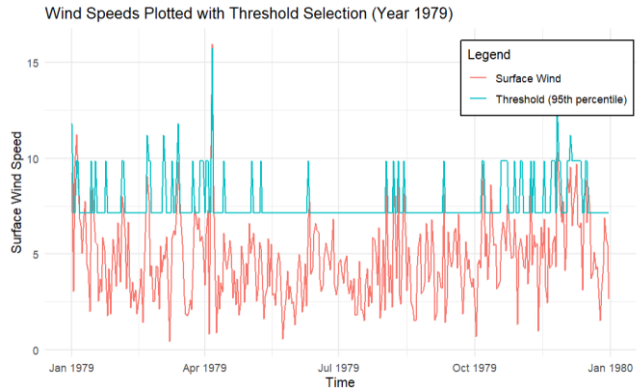
- We collect 1000, 850, and 500hPa wind speeds and temperature from a 10x10 grid centered around Toronto Pearson from ERA5 reanalysis dataset
- Calculate thermal wind (temperature gradients)
  - Take temperature 5 degrees north and 5 degrees east and calculate temperature difference from Toronto Pearson for temperature gradient
  - Thermal wind in x and y directions
- Calculate Atmospheric Stability
  - In layer 1(1000-850hPa) and layer 2(850-500hPa)
  - Caveat: only considers dry air, no water loading or humidity considered in calculation

Results

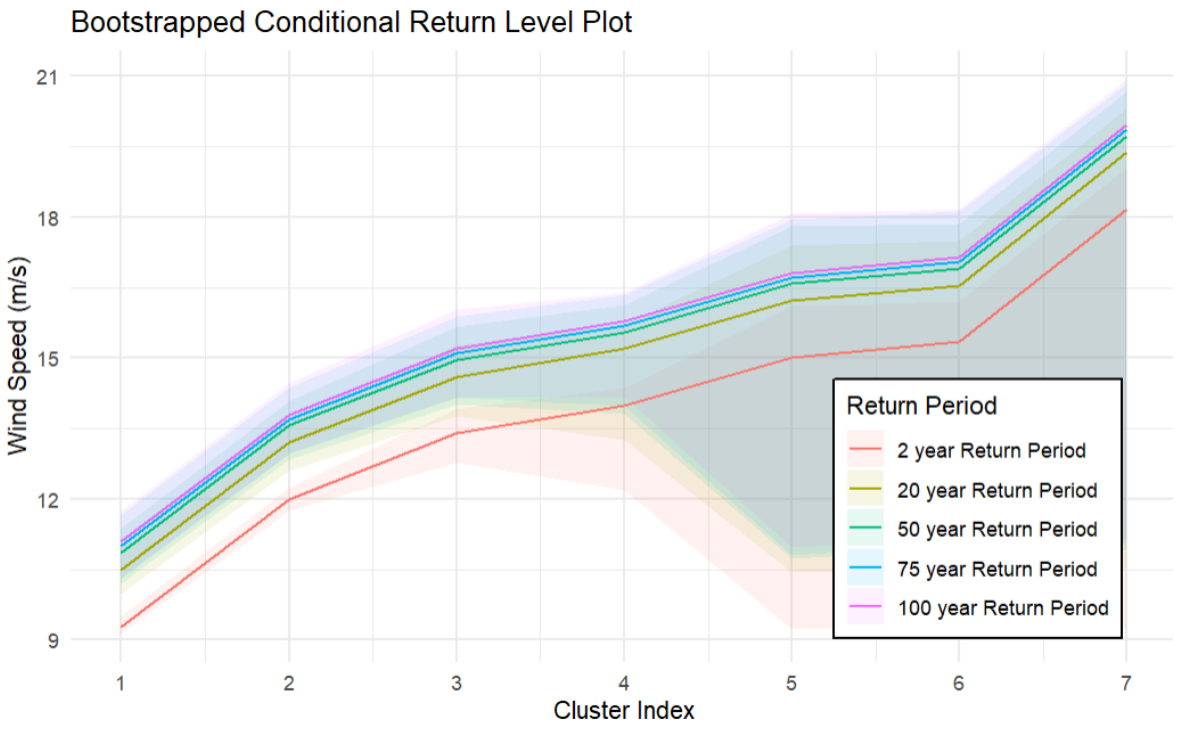
Parameter	Value	Std. Error
μ_0	9.087	0.066
μ_1	2.712	0.083
μ_2	4.106	0.29
μ_3	4.695	0.41
μ_4	5.723	0.593
μ_5	6.049	0.48
μ_6	8.875	0.862
σ	0.564	0.029
ξ	-0.117	0.024

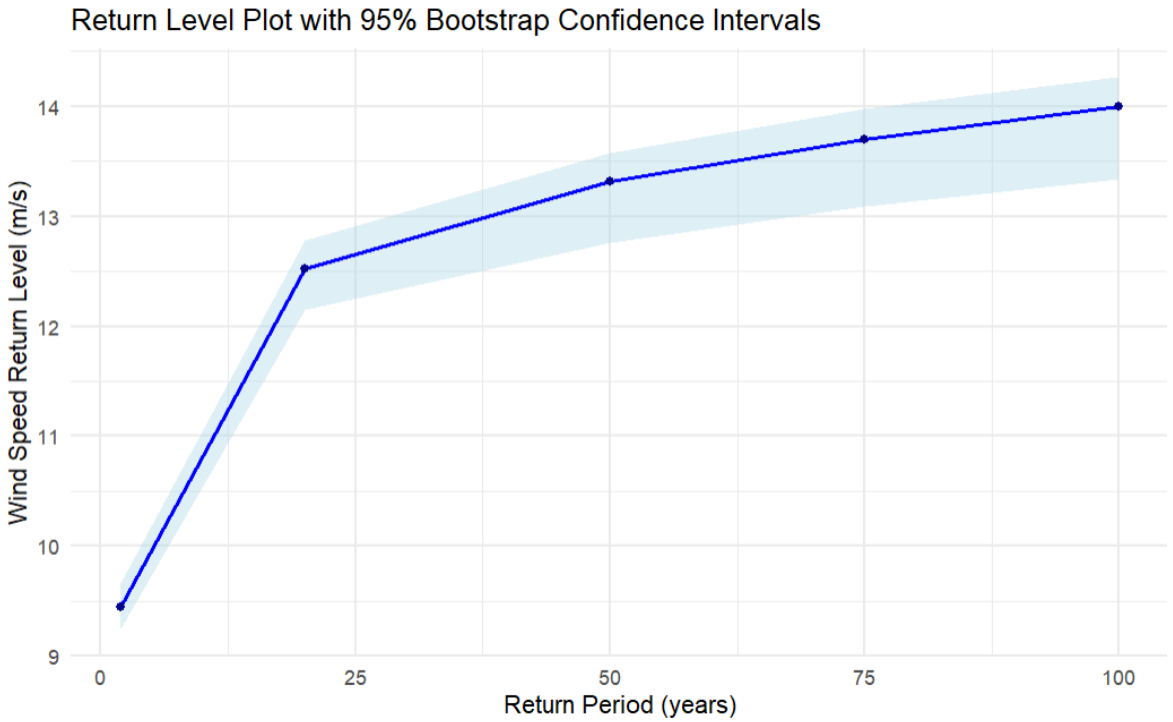




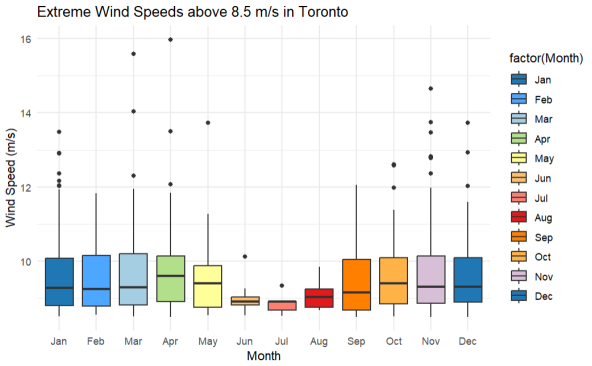


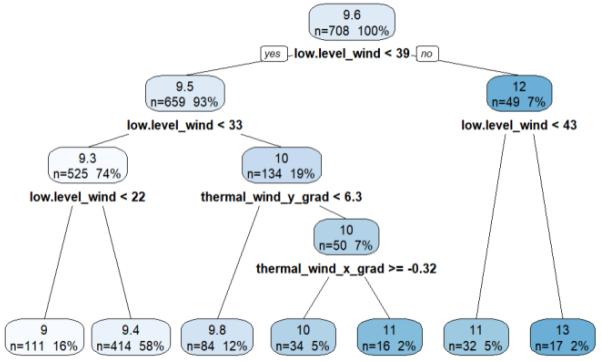
Return Level Plots





Problems



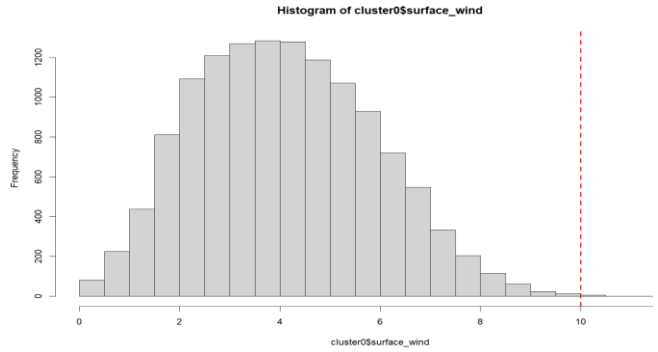


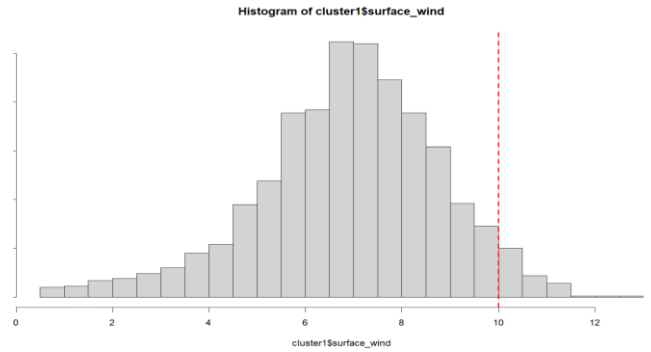
- Return level for 100 year period is only 14 m/s, not as high as 16 m/s from summer index model
  - Yet, QQ-plot suggests the model is a better fit
- Theory: Clustering nearly perfectly captured all possible extremes
  - Proportions of clusters with many extremes are very small, thus very little impact from them in the calculation

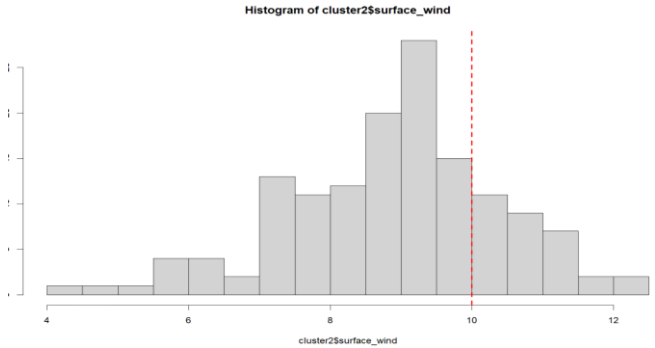
$$\mathbb{P}(Z > z_p) = \sum_{k=1}^{m} \mathbb{P}(Z > z_p | Z \in C_k)\mathbb{P}(Z \in C_k) = p$$

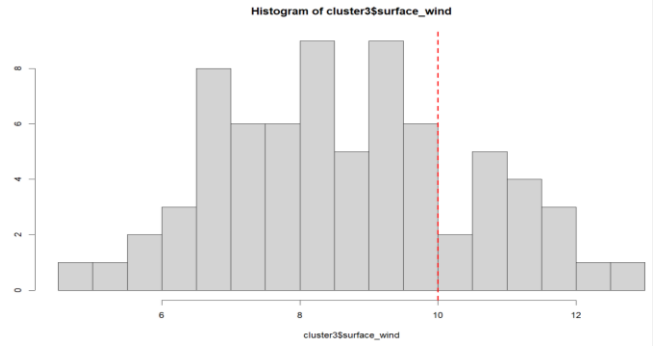
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Proportion	0.84	0.14	0.0093	0.0047	0.0016	0.0027	0.0008

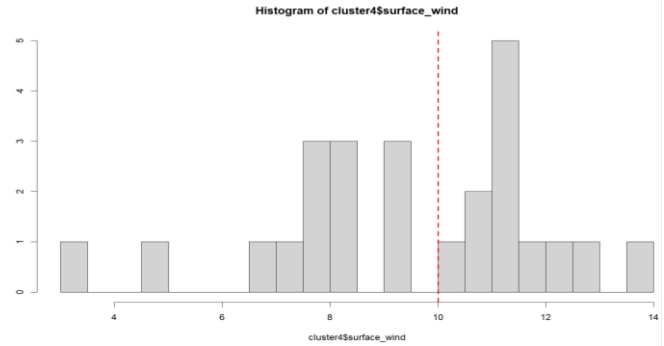
Distribution of each cluster

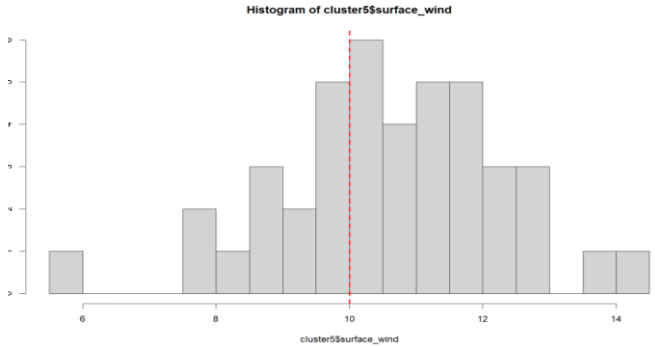


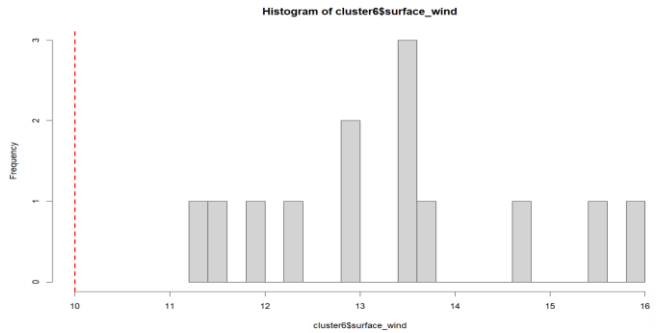












Future Work

- Provide simulations that correspond to a similar situation here to see if such return level calculation is realistic
- Try to use negative log-likelihood objective function and compare results with ANOVA

References

[1] Stuart Coles. An Introduction to Statistical Modeling of Extreme Values. Springer London, 2001.

[2] Jonathan A. Tawn Emma F. Eastoe. Modelling non-stationary extremes with application to surface level ozone. Journal of the Royal Statistical Society. Series C (Applied Statistics), 58(1):25–45, 2009.

[3] Sebastien Farkas et al. Generalized pareto regression trees for extreme event analysis. Journal of Extremes, 27:437–477, 2024.

[4] Lorenzo Mentaschi et al. Non-stationary extreme value analysis: a simplified approach for earth science applications. Hydrology and Earth System Sciences Discussions, 127:353–369, 2016.