Non-Stationary Modeling of Extreme Events using Extreme Value Theory and Regression Trees

Morgan Huang, Paul Kushner, Karen Smith

August 2025

Abstract

In environmental applications, stationarity is more than often not a valid assumption in today's framework due to climate change. If stationarity can be assumed, then extremal data can be easily modeled using simple applications of extreme value theory. Unfortunately, there is no general theory in working with non-stationary processes, thus statisticians working especially in this field are presented with a non-trivial task. In this study, we present a modification of a methodology presented by (Farkas et al. 2024) [8], which combines a general theory of extremes along with regression trees. We then apply this methodology to two datasets - ERA5 reanalysis from 1979 to 2020 and VR-CESM, where we analyze climate simulations between the year 2000 and the year 2090, both of which analyzes extreme wind speeds in Toronto. We suggest that this methodology helps us accurately model non-stationary extremes, as well as provide us with useful covariate analysis, in which we can analyze better the uncertainty of extremes and where it comes from.

1 Introduction

Extreme Value Theory (EVT) provides a powerful tool for climate scientists in determining return values of several weather extremes such as heavy rainfall (Friederichs, 2010)[13] and wave heights (Wang et al., 2021)[1]. The main theorems and results present in EVT are applicable to any distribution, with some assumptions, which makes its usefulness and practicality to real-world applications analogous to the Central Limit Theorem. However, caveats are still present in this evolving field, in particular concerning the data assumptions required for the main results in EVT to hold.

Like the Central Limit Theorem, EVT requires the data to be independent and identically distributed, but this is not the case especially in environmental applications, where serial dependence and nonstationarity are common. In fact, concerns about climate change that significantly affect the statistical behavior of climate variables suggest that a stationary assumption is no longer valid for current and future climate (Milly et al., 2008)[2]. Therefore, there is a bigger need in developing modern theories in EVT that can adapt to nonstationary data as well as an analysis on not just historical data but future climate projections given by climate models as well for more accurate and informed analysis.

Several modern research studies aim to address such problems. To account for serial dependence in extremes, analysts commonly use declustering techniques such as runs declustering (Coles, 2001)[9] and intervals declustering (Ferro and Segers, 2003)[12]. However, declustering is often wasteful of data, as we only fit clusters instead of individual data points into the model, and such dependency may provide useful information in the analysis. An alternative is to instead assume a first-order Markov Chain, and modify the likelihood to incorporate bivariate distributions which can account for first-order dependency in extremes (Smith et al., 1997)[3]. Since there is limited software for first-order Markov Chains in EVT and daily wind extremes were shown to have low autocorrelation ($\rho < 0.2$), we safely assume time independent data.

For nonstationary data, statisticians may implement regression parameters on separate covariates to EVT distribution parameters so that EVT distributions may adapt to different circumstances. However, there are no stable and concrete ways of including regression parameters into the model, giving way to subjective analysis, which may not be reliable. Moreover, using continuous covariates in the distribution parameters makes the general return level calculation tricky. An alternative is to split the non-stationary data into quasi-stationary slices and perform stationary EVT on each slice (Vousdoukas et al., 2016)[7]. A drawback is that for each model fitting process, only a subsection of the data is included in the likelihood, and consequently making the analysis on each slice less precise. Finally, analysts may choose to convert the nonstationary data into stationary data by removing seasonality and long-term trends. (Eastoe and Tawn, 2009)[11] removes nonstationarity by fitting the data into a Box-Cox location-scale model, while (Mentaschi et al., 2016)[5] uses a technique similar to regular normalization techniques, where they subtract the trend and divide the varying amplitude of the target variable across time. The challenging part of this method is estimating the trends and seasonality that is being subtracted from the data, which, like the first method presented here, induces a level of subjectivity to the analysis.

In this paper, to address nonstationary wind extremes (wind extremes have been shown to have strong seasonal dependence), we use a regression tree, as introduced in the paper by (Farkas et al., 2024)[8] to separate our extreme data into discrete clusters with respect to our covariates. We then define binary variables that correspond to these clusters and fit them into our EVT model with regression parameters using a point process approach and a varying threshold that is dependent on the clusters. Such methodology allows us to use all the available data in the model fitting process, splits the nonstationary data into quasi-stationary clusters, and provides a much simpler joint return level calculation, as it only involves a simple summation rather than multiple integration. Moreover, it helps us to analyze the relationship between wind extremes and its covariates more thoroughly, since regression tree partitioning allows multiple covariates to interact with each other. Interaction effects in simple regression analysis can be hard to interpret at times, especially when multiple terms are interacting with each other, thus this approach is much more favorable.

We perform two different analyses here, each regarding wind extremes in the Toronto Pearson Airport region. We first analyze ERA5 reanalysis data between the years 1979-2020, then perform a historical (2000s) versus future (2090s) climate model analysis using a variable resolution global climate model (VR-CESM), where the 2090s simulations simulate the worst case climate forcing scenario (RCP8.5), as used by (Morris et al., 2024)[6]. Note that a comparison using regular climate models with grid cells of around 100 km in length will not show any significant increase in wind extremes in Southern Ontario, as Morris 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.

This paper will be split as follows. Section 3 will showcase the relevant statistical theory, with the main theorems of extreme value theory presented in 3.1 (Coles, 2001)[9] and regression trees in 3.2. Return levels, which are the end-goal of extreme value theory, will be presented in 3.3. Section 4 presents an application to extreme wind speeds using two datasets - a historical ERA5 dataset in 4.1, and a comparison in climate projections between historical and future data points using VR-CESM in 4.2. We discuss the results and caveats of our methodology in the discussion section in section 5.

2 Methodology

2.1 Extreme Value Theory

As previously mentioned, at the core of EVT is a powerful limit theorem that can be applied to any distribution, similar to the famous Central Limit Theorem. Just as CLT takes the sample mean from any distribution and it will converge to a normal distribution, in EVT we take the maximum of evenly separated blocks across time, and the maximas can be shown to follow a certain class of distributions called the Generalized Extreme Value(GEV) distribution.

2.1.1 Stationary Processes

Theorem 1 (Fisher-Tippett-Gnedenko) *Suppose $X_1, X_2, \ldots,$ are independent and identically distributed random variables. For large enough block size n, define $M_n = \max\{X_1, \ldots, X_n\}$ If there exists constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$\mathbb{P}(\frac{M_n - b_n}{a_n} \leq z) \to G(z)$$

for some non-degenerate distribution function $G(z)$, then G is a is a member of the GEV family, where

$$G(z) = \exp\{-[1 - \xi(\frac{z-\mu}{\sigma})]^{-\frac{1}{\xi}}\}$$

defined on $\{z : 1 - \xi(\frac{z-\mu}{\sigma}) > 0\}$, where $\mu \in (-\infty, \infty), \sigma > 0$, and $\xi \in (-\infty, \infty)$.

Note that the theorem can be generalized such that we do not need to know the constants $\{a_n > 0\}, \{b_n\}$. Data that follows the GEV distribution can be summarized by the three parameters μ, σ, and ξ, where they represent the location, scale, and shape parameters. A characteristic worth noting is the ξ parameter, where the parameter governs the tail behavior of the distribution. If $\xi < 0$, then the distribution of M_n has lighter tails, while $\xi > 0$ has heavier tails. We can showcase this property by plotting return level plots, where return levels are define as follows:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - (-\log(1-p))^{-\xi}], & \xi \neq 0 \\ \mu - \sigma \log(-\log(1-p)), & \xi = 0 \end{cases}$$

where $\mathbb{P}(M_n > z_p) = p$, and z_p refers to the return level corresponding with the return period $1/p$. In words, we interpret the return level z_p as the value expected to be reached or exceeded once every $1/p$ blocks (heuristically, we choose n to be the length of a year, so z_p refers to the value exceeded once every $1/p$ years). Return levels are often the end result of extreme value analysis, as it provides estimates of how extreme a certain random variable may get in a future time period, given that historic and future time periods are stationary.

We can create confidence intervals of return levels by assuming that the parameters follow a normal distribution, by MLE properties, and applying the delta method to $Var(z_p)$:

$$Var(z_p) = \nabla z_p^T V_{(\mu,\sigma,\xi)} \nabla z_p$$

where $V_{(\mu,\sigma,\xi)}$ is the variance-covariance matrix of the GEV parameters. However, due to the normality assumption of the MLE, this method often results in symmetric confidence intervals which may not reflect the underlying nature of the uncertainty(we may prefer skewed CIs). An alternative, more accurate approach is to use the profile likelihood, which does not rely on any assumptions on the distribution of the model parameters.

To apply the theorem into practice, we typically use Maximum Likelihood Estimation(MLE) to estimate the parameters (μ, σ, ξ), where we split the data into annual blocks, take its maximum, and use the maximas to fit the model, hence why this method is called the 'block maxima' approach. Model validation is also important in checking if the model assumptions we assumed are realistic, where we typically use QQ-plots, PP-plots, and return level plots to assess correctness. One can use quantitative methods in model checking such as the Lilliefors test(Wilks, 2018)[14], however it requires statistical simulations in finding its critical values, and statistical software typically only offer the statistical test for normal data.

It is worth nothing that the block maxima approach is often wasteful of data, as there are many instances where extremes lie in each block but are 'tossed away' since there are bigger extremes in its corresponding block. We thus provide an alternative approach to extreme value modelling:

Theorem 2 *Let $X_1, X_2, \ldots$ be i.i.d with a common CDF F. Let*

$$M_n = \max\{X_1, \ldots, X_n\}$$

and suppose that $\mathbb{P}(M_n \leq z) \to G(z)$, where G is a member of the GEV family. Then, for large enough u, the distribution function of $X - u$ conditional on $X > u$ is given by

$$H(y) = 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-1/\xi}$$

where $y = X - u$ for $X > u$, $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, and $\tilde{\sigma} = \sigma + \xi(u - \mu)$, where (μ, σ, ξ) are GEV parameters.

H is called the Generalized Pareto distribution, and is dependent on a hyperparameter u, which represents the threshold of our model. For a given u, we say that a data point is considered 'extreme' if $X > u$. We see that the choice of u corresponds to a bias-variance tradeoff, where a u that is too large induces large variance in the model, since there is likely not enough data fitted into the model. In addition, u that is too small induces model bias as we are likely to violate model assumptions. To choose a suitable threshold u, we rely on GP distributional properties. Notably, if $X - u|X > u$ is GP-distributed, then it follows that

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

for $\xi < 1$. Then, the limit result should still hold for a larger threshold $u_0 > u$, and thus $\mathbb{E}(X - u_0|X \geq u_0) = \frac{\sigma_{u_0}}{1-\xi}$, since the threshold choice should not affect ξ. Thus, we can rewrite the equation as:

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

Therefore, $\mathbb{E}(X - u|X > u)$ is a linear function of u. Using a mean residual life plot from our empirical data:

$$\{(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)} - u))|u < x_{\max}\}$$

where $x_{(i)}$ are the n_u threshold excesses above u and $\frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)}-u)$ is the empirical average of the threshold excesses, we choose the smallest u such that our mean residual life plot is approximately linear on u, by the GP property mentioned above.

Another way to choose a threshold is to reparameterize the shape parameter as $\sigma^* = \sigma_{u_0} - \xi u$, and plot both σ^* and ξ with respect to u. It can be shown that the choice of u does not affect the values of σ^* or ξ if model assumptions are true, thus we choose the lowest u such that the plots of u against σ^* and ξ are approximately constant.

Regardless of the above methods available in choosing a threshold, there is still no novel way of choosing a threshold, as it is possible that the above properties are true for all different threshold values u, even though many choices lead to model bias(it is not a unique property).

2.1.2 Non-stationary Processes

As mentioned before there are many cases in which data contains trends and seasonality, and thus the assumption of stationarity is no longer realistic. Extremes at one time point may not be considered extreme at another time point. An easy, although subjective, workaround is to include regression parameters into the model(Davison and Smith 1990)[10]. This allows us to model a different extreme-valued distribution for each time step t and model trends and seasonality. For instance, we can model a random variable Z_t from the block maxima approach as:

$$Z_t \sim GEV(\mu(t), \sigma(t), \xi(t))$$

where $\theta(t) = h(X^T\beta)$, where θ is a GEV parameter, h is some function of time, X represents covariates that depend on time t, and β is a vector of regression parameters. Choices of regression models inside each parameter include a linear model, a polynomial, or cyclic functions, where the parameter is a Fourier Series. We typically choose h to be an indicator function for the location parameter and the exponential function for the scale parameter to ensure positive σ, and ξ is to be assumed constant over time.

It is common for scientists to compare whether or not a non-stationary model or a stationary model is sufficient for the data at hand. To test whether or not a stationary model generalizes the data better than a nonstationary model, we can use the likelihood-ratio test, since the stationary model with no regression parameters is a simplified model compared with the non-stationary model. The deviance statistic is given by:

$$D = 2(l_1(M_1) - l_0(M_0)) \sim \chi^2_k$$

where k is the degrees of freedom, which is equal to the difference in the number of parameters in M_0 and M_1.

Since for each time-step, the distribution of extremes will be different, when performing model diagnostics we must standardize each distribution before drawing out the QQ/PP plot. For block-maxima models, we typically standardize them into a standard Gumbel distribution, while for threshold models we standardize them into a standard exponential distribution, which is a special case of the GP distribution.

For threshold models, we can also realistically implement a varying threshold which changes with respect to time, $u(t)$, especially when linear or cyclic trends occur in the data. However, problems arise since threshold stability properties that are tied to the threshold model become violated(Eastoe and Tawn, 2008)[11]. In particular, the following does not hold:

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

for any $u_0 > u$, and covariates X. A workaround is to use the point process (PP) approach which combines the use of a threshold from POT models and GEV parameters from the block maxima approach. This enables us to use a varying threshold since threshold stability does not exist in the PP approach as we are using GEV parameters instead of the GP parameters.

Theorem 3 *Let $Z_1, Z_2, \dots$ be i.i.d random variables with $M_n = \max\{Z_1, \dots, Z_n\}$, for some large n, where there exists $\{b_n\}$ and $\{a_n > 0\}$ such that $\mathbb{P}((M_n - b_n)/a_n \leq z) \to G(z)$, where G is part of the family of GEV distributions. Let z_- and z_+ be the lower and upper end points of G. Then, for any threshold $u > z_-$, the sequence of point processes:*

$$N_n = \{(i/(n+1), (Z_i - b_n)/a_n) : i = 1, \dots n\}$$

converges in the region of $[0,1] \times (u, \infty)$ to a Poisson process with intensity measure on set $A = [t_1, t_2] \times (z_-, z_+)$, given by:

$$\Lambda(A) = (t_2 - t_1)[1 + \xi(\frac{z - \mu}{\sigma})]^{-1/\xi}$$

where (μ, σ, ξ) are the location, scale, and shape parameters of the GEV distribution.

A simple explanation for this result is that after defining some suitable threshold u, we can derive the probability of an arbitrary point in set N_n being above the threshold u as a function of the GEV parameters:

$$p = \mathbb{P}((Z_i - b_n)/a_n > u) \approx \frac{1}{n}[1 + \xi(\frac{z - \mu}{\sigma})]^{-1/\xi}$$

The number of points in N_n that are above threshold u can be modeled using a Binomial distribution $N_n \sim Bin(n, p)$, and thus converges to a Poisson process with intensity measure Λ, by Binomial convergence results.

2.1.3 Return Levels

As we are working with nonstationary extremes, return levels are much harder to calculate, unless we are conditioning on a certain covariate. For example, say our PP model depends on some covariate X. Then, we can calculate conditional return levels by solving for z_p from the equation

$$\mathbb{P}(Z > z_p | X = x) = p$$

If we are interested in the general behavior of the extremes of Z, then we must integrate out X and solve for z_p, as follows:

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

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

$$\mathbb{P}(Z > z_p | X = x) = 1 - \exp\left\{-1\left[1 - \xi(x)\Big(\frac{z_p - \mu(x)}{\sigma(x)}\Big)\right]^{-1/\xi(x)}\right\}$$

If we include multiple covariates $X_1, X_2, \dots$, then we require multiple integrals to solve for z_p, which is not ideal. Using clusters derived from regression trees, we partition the data into finite clusters with respect to the covariates $X_1, X_2, \dots$, and thus we only have to calculate the conditional return level conditioned on each cluster, and solve for z_p using a summation. If we have clusters $C_1, \dots, C_m$, for m clusters, then we have:

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

2.2 Regression Trees

We have two goals that we want to achieve through clustering. From a modeling perspective, since our data is nonstationary, we want to find clusters such that each cluster is approximately stationary. In addition, clusters should be differentiable from each other such that we can understand better where the uncertainty of wind extremes come from, which is why we want our clusters to be defined based on our covariates. We propose a regression tree algorithm the CART (Clustering and Regression Tree) algorithm (Breiman et al. 1984)[4].

Regression trees are created by iteratively splitting a root node into two branches using a set of rules that are dependent on the covariates, where a certain objective function is minimized. The objective function is defined as:

$$\theta^*(x) = \arg\min_{\theta \in \Theta} \mathbb{E}\Big[\phi(Z_i, \theta)\Big| X = x\Big]$$

where ϕ is some loss function, and $\theta \in \Theta$ is the parameter space. In this case, Θ will be the space of the GEV parameters. The idea is to find a certain split of the covariates such that the total loss function of the two child nodes T_1 and T_2 is less than the loss of the root node T_0:

$$\phi(Z, \theta | Z \in T_0) > \Big[\phi(Z, \theta | Z \in T_1) + \phi(Z, \theta | Z \in T_2)\Big]$$

Splitting the tree based on covariates relies on the use of a set of rules. An example of a rule is $R_l(x) = \mathbb{1}_{x_1 < x < x_2}$, for some $x_1, x_2 \in dom(X)$. At each split, new rules are created and are more 'specific' then its predecessors. More precisely, for some rule $R_l(x)$ and split k where we split the node corresponding to rule $R_l(x)$, the following must be true:

$$\{x | R_{l+1}(x) = 1\} \subset \{x | R_l(x) = 1\} \qquad \{x | R_{l+2}(x) = 1\} \subset \{x | R_l(x) = 1\}$$

where rule $R_l(x)$ is split into two rules $R_{l+1}(x)$ and $R_{l+2}(x)$. In addition, we must also have

$$\{x | R_{l+1}(x) = 1\} \cup \{x | R_{l+2}(x) = 1\} = \{x | R_l(x) = 1\}$$

and

$$R_{l+1}(x) R_{l+2}(x) = 0$$

for all x. Once we fully split the tree, we obtain a set of rules $(R_1(x), \ldots, R_n(x))$. The set of rules must satisfy the following properties:

1. $\forall x \in dom(X) \subset \mathbb{R}^d, R_l(x) = 1$ or 0, depending on some conditions of the covariates on rule l

2. At split k with rules $(R_1, \ldots R_{n_k})$, $R_{l_i}(x) R_{l_j}(x) = 0$, for $l_i \neq l_j$

3. $\sum_l R_l(x) = 1$

To derive the set of rules $R_l(x)$, we use a grid search algorithm over all possible covariates and their values, then choose the covariate and its covariate value such that the loss function is minimized over all possible splits. If a suitable split cannot be found; that is, if the condition $\phi(Z, \theta | Z \in T_0) > \Big[\phi(Z, \theta | Z \in T_1) + \phi(Z, \theta | Z \in T_2)\Big]$ cannot be satisfied, then we stop the growth of the tree. Once we develop the maximal tree, we prune the tree to find the splits that minimize the objective function the most using cross-validation, but this step will be omitted from the analysis and the number of clusters we want from our regression tree will be fixed arbitrarily at 7.

This summarizes the CART algorithm, for some arbitrary loss function ϕ. A common loss function is the ANOVA method, where the loss function is the squared deviance:

$$\phi(Z, m(X)) = \sum_{Z \in \mathcal{T}_l} (z - m(x))^2$$

where $m(x) = \mathbb{E}[Z|Z \in \mathcal{T}_l]$, and $\mathcal{T}_l$ represents data corresponding to the lth leaf . (Farkas et al. 2024)[8] proposes the negative log-likelihood function as the loss function instead, where he uses a threshold model with GP parameters. Compared with ANOVA, the negative log-likelihood loss function has a stronger theoretical basis, as each split the regression tree performs signifies that the algorithm has found a partition of the dataset in which the two branches are 'more stationary' than its root. Moreover, ANOVA prefers closer points than further points in each cluster, and therefore does not acknowledge that far away, more extreme, points can belong to the same stationary process.

Unfortunately, the RPART package which implements regression trees cannot support negative loss functions, which the negative log-likelihood is capable of violating. Farkas adds arbitrarily large values for each loss function that depends on the number of data points in its corresponding leaf, but it is very arbitrary and hard to find suitable large values. Thus, we are forced to use the ANOVA method.

2.3 Fitting Process

For both datasets, we analyze daily maximum Toronto Pearson surface wind speeds at 1000hPa, using the following atmospheric covariates: low level wind speeds at 850hPa, jet stream winds at 500hPa, thermal wind in the x and y direction, and atmospheric stability in 2 layers (1000-850hPa and 850hPa-500hPa). Note that thermal wind is calculated by taking the temperature difference between Toronto Pearson and the temperature 5 degrees north of Toronto Pearson, as well as the temperature 5 degrees east of Toronto Pearson. In addition, we assume dry air in our analysis.

We will first fit our data into our regression tree using the ANOVA method and prune the tree such that we obtain 7 leafs representing 7 clusters. Then, we use those clusters to define binary variables that represent each cluster. More generally, if we have m clusters, for some $m \in \mathbb{Z}$, then we define $m-1$ binary variables corresponding to each cluster, leaving one cluster acting as a dummy variable:

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

Then, we define our GEV parameters using the binary variables as follows:

$$\begin{aligned} \mu(\vec{X}) &= \mu_0 + \mu_1 X_1 + \cdots + \mu_{m-1} X_{m-1} \\ \sigma(\vec{X}) &= \exp(\phi_0 + \phi_1 X_1 + \cdots + \phi_{m-1} X_{m-1}) \\ \xi(\vec{X}) &= \xi_0 + \xi_1 X_1 + \cdots + \xi_{m-1} X_{m-1} \end{aligned}$$

In our model, we will also include a varying threshold, as we assume that each cluster represents a different stationary process with different 'definitions' of extreme values. For simplicity, we heuristically choose the 95th percentile surface winds at each cluster as the cluster threshold, where our threshold for our model is defined as:

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

where u_{k+1} represents the threshold for the kth cluster, for $k = 1, \ldots, m-1$, with u_0 as the threshold for the cluster corresponding to our dummy variable.

Then, we fit our point process nonstationary model with regression parameters and a varying threshold using maximum likelihood estimation to obtain estimates for $\vec{\mu}, \vec{\phi}$, and $\vec{\xi}$. We fit multiple models by including and omitting some parameters and select a suitable model using likelihood ratio tests. QQ-plots will be plotted to validate our model. We then calculate conditional return levels for each cluster and calculate the joint return levels using the empirical proportions of clusters in the data $\mathbb{P}(Z \in C_k)$. A simple bootstrap algorithm using our selected model will be performed to obtain confidence intervals for our conditional and joint return levels.

3 Results

3.1 ERA5 Dataset

Recall that we analyze the surface wind speeds between the years 1979-2020. We provide a short summary of our data to showcase the seasonality of our data.

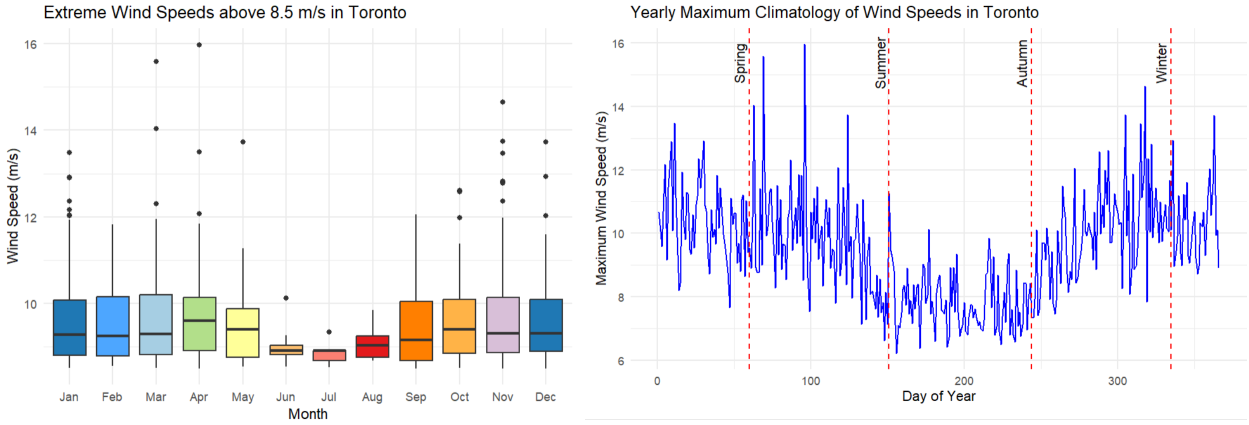
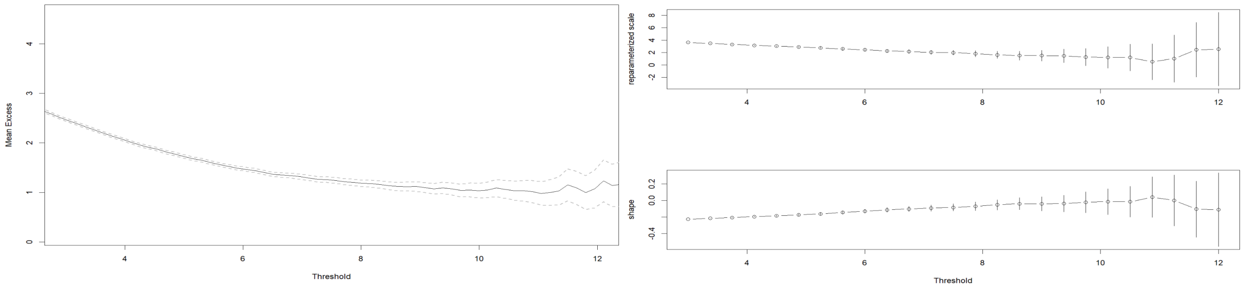


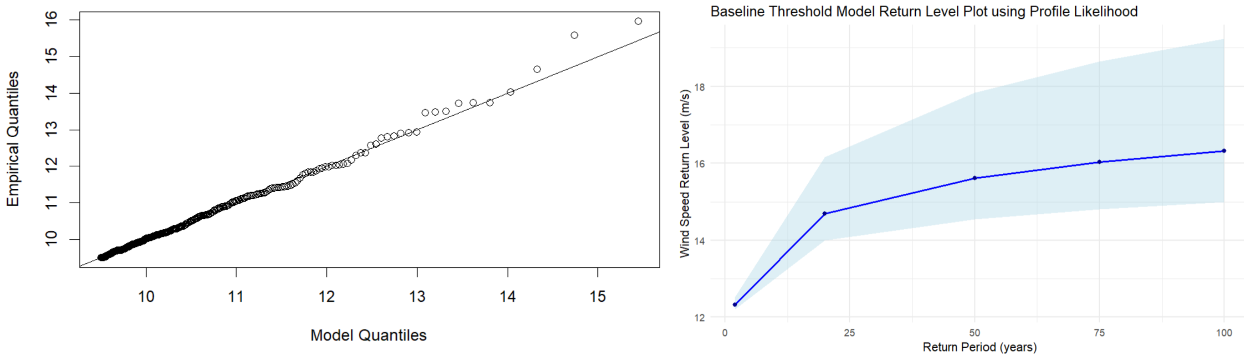
Figure 1: *Boxplots (left) showcase the different distribution of extreme wind speeds in the summer months compared with the non-summer months. Daily Maximum Climatology (right) is plotted by taking the maximum of each day of year throughout the dataset.*

As mentioned beforehand, seasonality is a great indicator of our data being non-stationary, thus motivating our proposed methodology. Visually, we can split our data into summer and non-summer months, and try to assume that both slices are approximately stationary. However, there may be some underlying structure in the data that may show non-stationarity given some combination of covariates in our data, which motivates the use of regression trees. As a reference model, we create a baseline model that assumes all the data belongs to one stationary process.



a)

b)



c) d)

Figure 2: *a) Mean Residual Life Plot, b) Threshold Stability Plots, c) QQ-Plot, d) Return Level Plot*

From a) and b), we can approximate that the ideal threshold for our baseline model is around 8 m/s, since that is the point where the mean residual life plot stops curving and is, afterwards, approximately linear, and when the reparameterized scale σ^* and ξ are approximately constant. After fitting our point process model using the 8 m/s threshold, we get a QQ-plot in c) that shows that our model does fit well, except at the tails, where it slightly underestimates the extreme tails. In d), using profile likelihood methods we get the return level plot as a function of the return period with its corresponding 95% confidence interval. As an example, the 100-year return period corresponds to a return level of around 16 m/s, with a 95% confidence interval of (15, 19.24).

We can also create a model using the summer index split, which gives us a better fit than the reference model, with a p-value of $< 2.2\text{e}{-16}$ using the likelihood ratio test. However, for brevity sake, we will skip this step in this paper and model extreme winds using a regression tree, which also gives a statistically significant likelihood ratio result compared with the summer index split (also $< 2.2\text{e}{-16}$). To construct the regression tree, we note that we only want to characterize extreme wind speeds, thus we filter the data to surface wind speeds of above 8.5 m/s. Such a value is chosen arbitrarily here, but the reader must note that different values can correspond to better or worse model fits. After growing and pruning the tree to produce 7 distinct clusters of the data, we get the following regression tree:

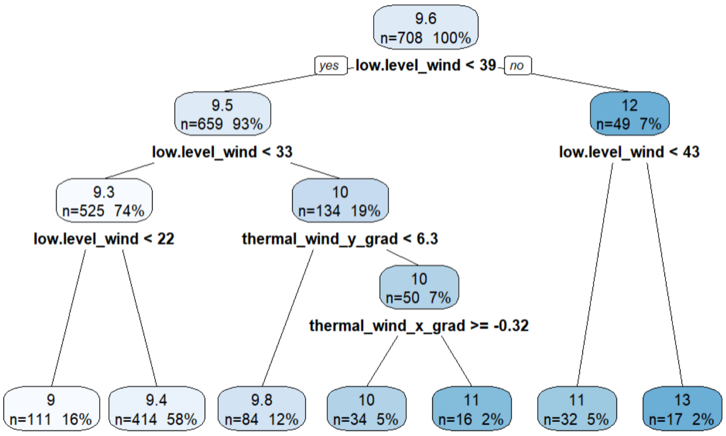


Figure 3: *Regression Tree produced using* ANOVA *from the* RPART *package. Covariate used and its split value is shown at every parent leaf. Each leaf of the tree contains the number of observations in the leaf along with the percentage of the total observations. The top number of each leaf denotes the mean cluster surface wind speed.*

Once we obtain the clusters, we theorize that each cluster has different 'definitions' of extreme value, thus we define a varying threshold using the 95th percentile cluster surface winds. We choose to rely on heuristics at this step, since the threshold stability plots and mean residual life plots at each cluster are very hard to interpret.

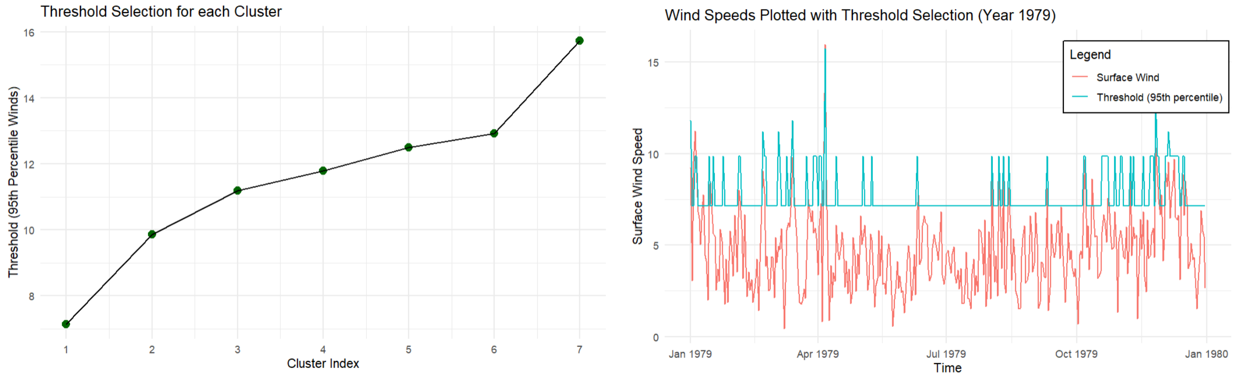


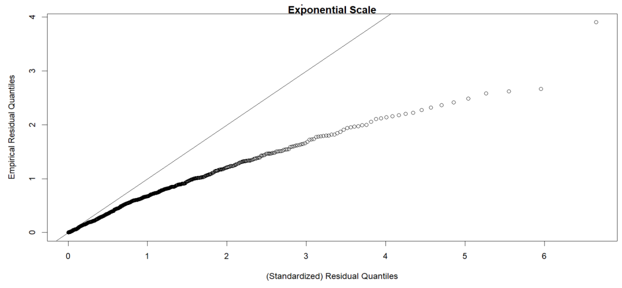
Figure 4: *Threshold Selection plot (left) for each cluster, defined as 95th percentile cluster surface wind speeds. The cluster index refers to the clusters defined in the regression tree, where an index of 1 represents the cluster at the leftmost leaf, and an index of 7 represents the cluster at the rightmost leaf. We also plot the varying threshold with empirical surface wind speeds (right) on the first year of the dataset (1979).*

Notice that we used surface wind speeds of above 8.5 m/s to construct our regression tree, however we use all the data to define our varying thresholds, which explains why the threshold of our first cluster is actually lower than 8.5 m/s.

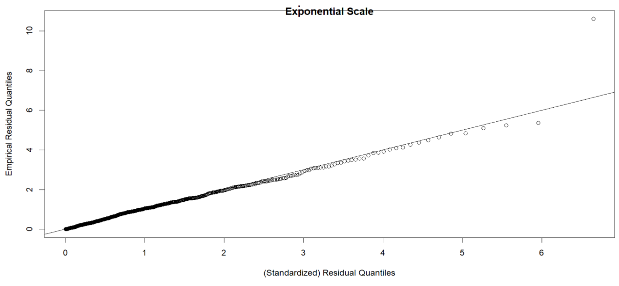
Using the clusters obtained in the regression tree as and a varying threshold, we then fit multiple models, each with different parameters, starting with just (μ, σ, ξ), then introducing regression parameters for each extreme-valued parameter one-by-one. We use the likelihood ratio test to select a suitable model.

Model	p	l	p-value
1. Varying Threshold	3	677.38	
2. As 1. but μ is linear with clusters	9	970.51	$< 2.2\text{e}{-16}$
3. As 2. but $\log \sigma$ is linear with clusters	15	975.81	0.1
4. As 2. but ξ is linear with clusters	15	975.1	0.1636

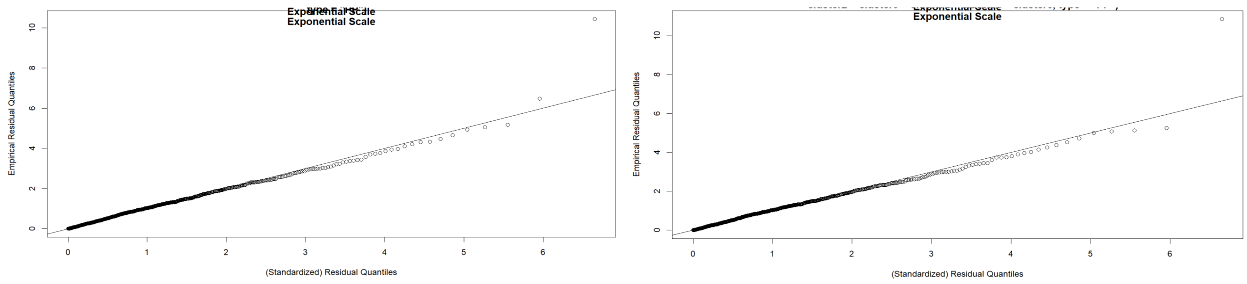
Table 1: *Model Selection results, where we choose the 2. model of using regression parameters at only the μ parameter by the likelihood ratio test. p denotes the number of parameters, and l denotes the likelihood of the model*



1.



2.



3. 4.

Figure 5: *QQ-plots of all 4 models, standardized to a standard exponential distribution. We can see that the first model does a poor job, with the second one being the best model (according to likelihood ratio test). The other two seem to do equally well compared with the second, thus its added complexity is deemed unnecessary. Notice the outlier of a wind speed of 16 m/s that failed to be modeled in each of the 4 models, compared to the reference model which did a better job at modeling such extreme value.*

The best model according to the likelihood ratio test and the QQ-plots is the second model, with regression parameters only in μ. Notice that the QQ-plot looks much better than the QQ-plot of our reference model except for one point, highlighting the modeling difference between a stationary and nonstationary assumption in the data. The parameter values and its standard errors are calculated as follows:

Parameter	Value	Standard 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

Table 2: *Table of Parameter Values for the second model*

μ_0 denotes the location parameter conditioned on cluster of index 1, whereas $\mu_0 + \mu_i$, for $i = 1, \ldots, 6$ denotes the location parameter conditioned on cluster of index $i + 1$.

We calculate conditional return levels and joint return levels in Figure 6. For its 95% confidence intervals, we use a bootstrap sampling algorithm 1000 times for each return level plot.

For the conditional return level plot, we see that the confidence intervals drastically grow larger as we move to the rightmost cluster of the regression tree. Since we used all the data in the model fitting process, the percentages of total observations in the regression tree are misleading. After de-filtering our data, we derive the true proportions of the clusters in the data in the following table:

	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

Table 3: *Table of proportions of regression tree clusters in the data. Cluster 1 represents 84% of the total observations, thus the conditional return levels are more precise and have more influence in the general return level calculation than the other clusters.*

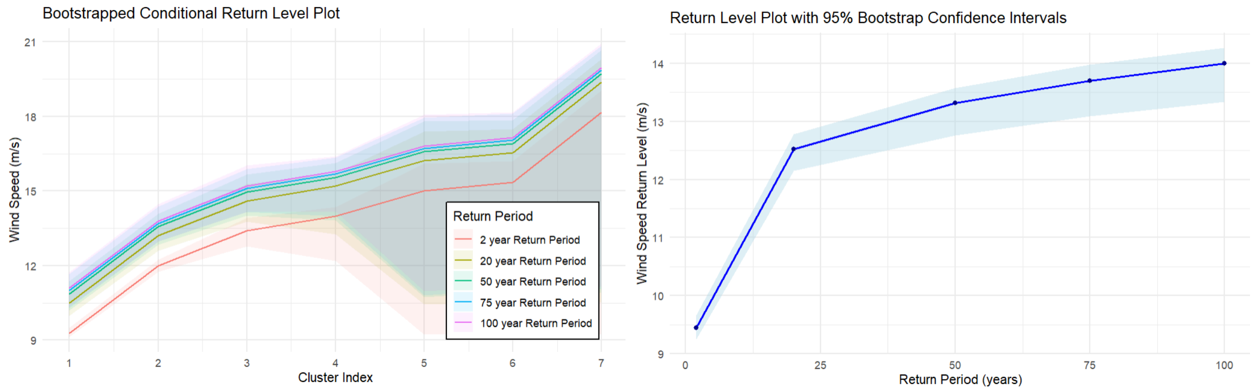


Figure 6: *Conditional Return Level Plot (left) and Joint Return Level Plot (right), each with 95% bootstrap confidence intervals*

A key difference between the regression tree model and the reference model is that the return levels here are much smaller than the baseline model, where, for instance, the 100-year return level is around 14 m/s for the regression tree model, and the return level with the same return period is around 16 m/s for the baseline model. In addition, when looking at the boxplots closely we can see that there are 4 data points that have a surface wind speed of above 14 m/s in a 42 year period. Thus, it seems unrealistic to have 14 m/s as the 100-year return level.

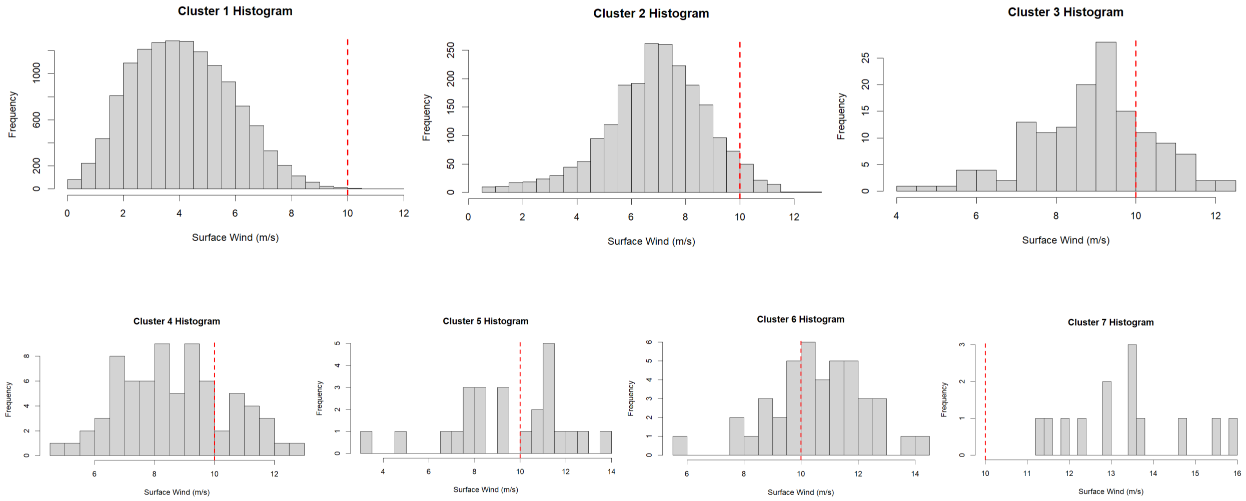


Figure 7: *Grid of Histograms of Surface Wind Speeds for each cluster. Red dotted line at each histogram acts as a reference point of 10 m/s, which showcases the proportion of extreme values in each cluster. Cluster 1 has the least proportion of extreme wind speeds, and as the cluster index increases, the proportion increases as well, up to cluster 7, whose observations are all above 10 m/s.*

A possible explanation for this phenomenon is to look at the proportion of the clusters $\mathbb{P}(Z \in C_k)$ in the joint return level equation. Since cluster 1 represents 84% of the total data, which means that the parameters corresponding to cluster 1 will have the greatest effect in the return level calculation. Analyzing the histogram (Fig. 7) and the parameter values (μ_0, σ, ξ) (Table 2), we can see that there are not many extreme values in the cluster, and its location parameter is relatively small compared to the other cluster location parameters. Thus, when calculating the general return level, the clusters with the greatest proportion of extreme values are overshadowed by the 1st and 2nd clusters, which have the smallest proportion and lowest location parameters.

3.2 VR-CESM Dataset

We would like to now turn our attention to climate models with historical and future projections of surface wind speed to access whether or not there is a significant climate change signal. We simulate daily wind speed and temperature climate from the year 2000 and 2090 around 30 times each, totaling around 10000 data points for each year. The atmospheric covariates will be exactly the same as in the previous case study. Note that the climate model has not been downscaled, thus the historical climate projections may look different than the ERA5 historical dataset.

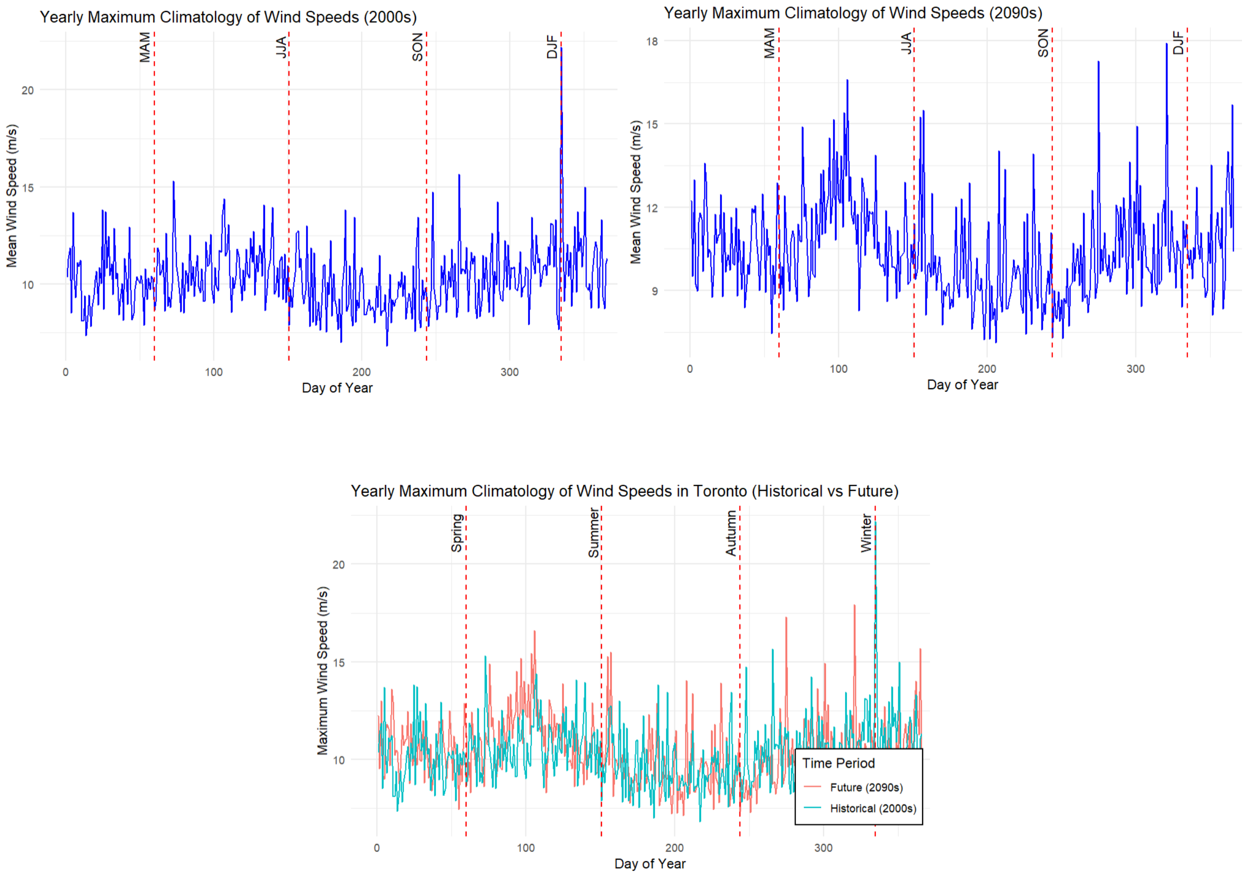


Figure 8: *Maximum Climatologies of each simulation, where (top left) corresponds to the 2000s simulation, (top right) corresponds to the 2090s simulation, and (bottom) shows a direct comparison between the two climatologies.*

As shown by the maximum climatologies above, there appears to be no real significance of an increase in extreme values between historical and future simulations. In addition, the seasonal cycle, which was visible in the ERA5 dataset, does not appear in the climate models. Only a slight increase in extreme winds can be seen in the fall and spring months here. We can look further into the extreme behavior more closely using a simple boxplot analysis, and use ANOVA and Wilcoxon signed-rank tests (parametric and non-parametric location tests, respectively) to test for a significant difference in mean extreme wind speeds for each month. Note that the alternative hypothesis for the Wilcoxon signed-rank test is a positive increase in wind speeds between the 2000s and the 2090s, whereas ANOVA only tests for a change in mean. In addition, we can also check if there were any significant decreases in wind speeds; it turns out that November is the only month with such behavior.

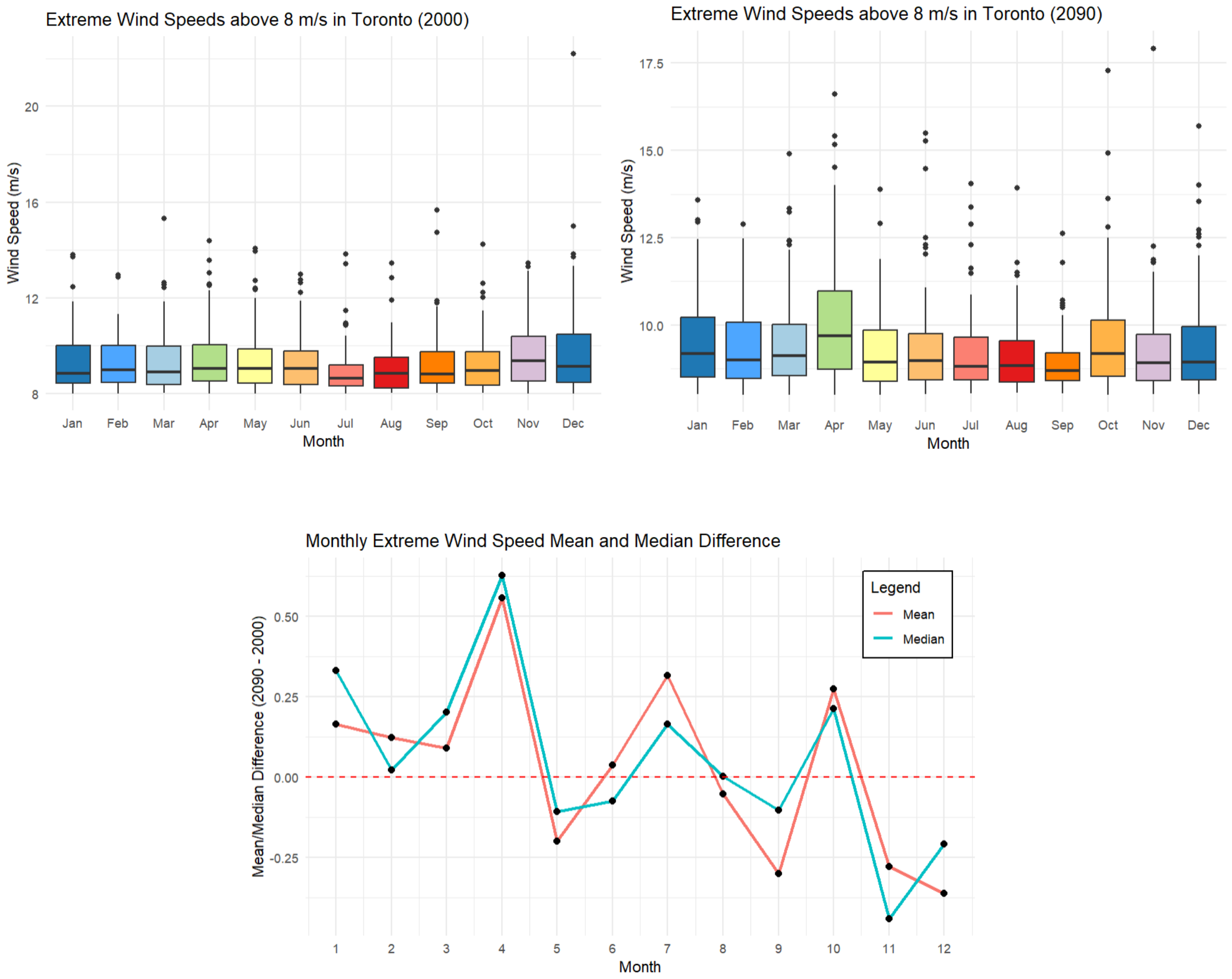


Figure 9: *Boxplots (top left shows 2000s and top right shows 2090s simulations) showcasing the weak seasonal cycle in both simulations. A key difference can be seen in the months of April and October, where the extreme wind speeds increase slightly in the 2090s compared with the 2000s. Differences in mean and median for each month of surface winds speeds can be shown at the bottom, with points above the dotted red line denoting an increase in 2090s wind speeds compared with 2000s wind speeds.*

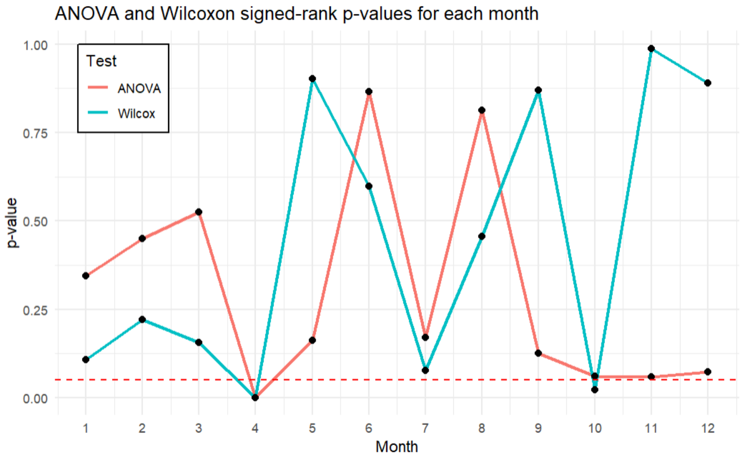


Figure 10: *P-values of ANOVA and Wilcoxon signed-rank tests for each month. We see that there is a significant increase ($p < 0.05$) in extreme wind speeds above 8 m/s in the months of April and October.*

As a reference model, we assume stationarity for both climate simulations, and fit threshold models for each simulation and compare its return levels. We check if historical and future surface winds change due to climate change.

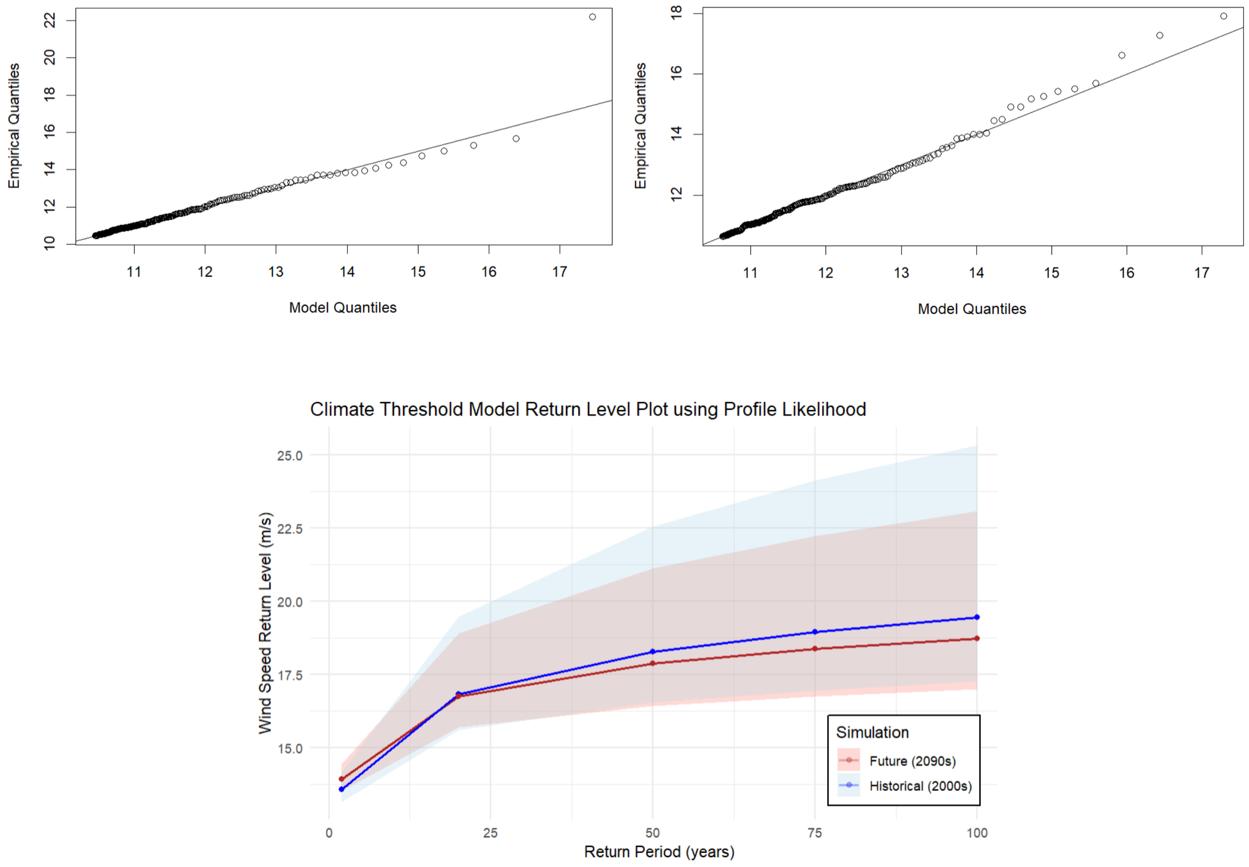


Figure 11: *QQ-plots of 2000s simulation (top left) and 2090s simulation (top right). Return level plot of both simulations (bottom) indicate a slight decrease in return levels.*

The above results show that there is no increase in extreme wind speeds, despite (Morris et al. 2024)[6] showing that there is a 3-5% significant increase in the 50-year return period of extreme wind speeds using the VR-CESM model. The extreme threshold he used is the 98th percentile, which we used here as well for both simulations, indicating contradictory results. We now repeat the same analysis using regression trees.

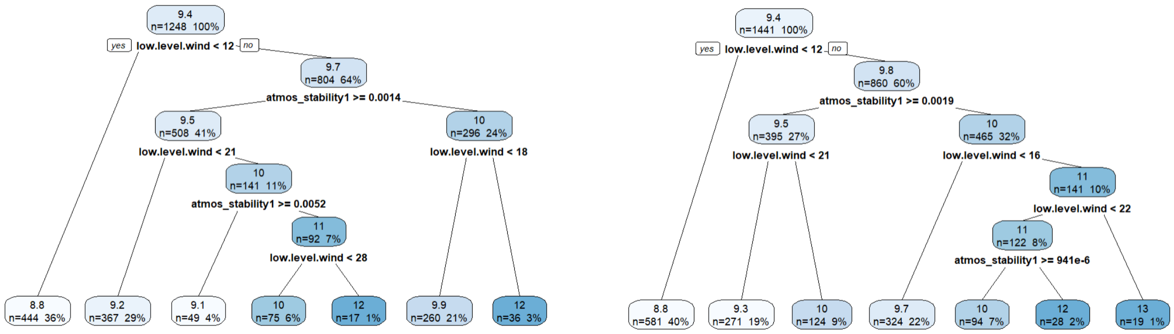


Figure 12: *Regression trees of the 2000s simulation (left) and 2090s simulation (right). We filter the data to only include wind speeds of above 8 m/s to construct the regression tree. Notice that the covariates used in the partitioning is different from the regression tree in the ERA5 reanalysis data, where the atmospheric stability of the 1st layer is more dominant than thermal wind.*

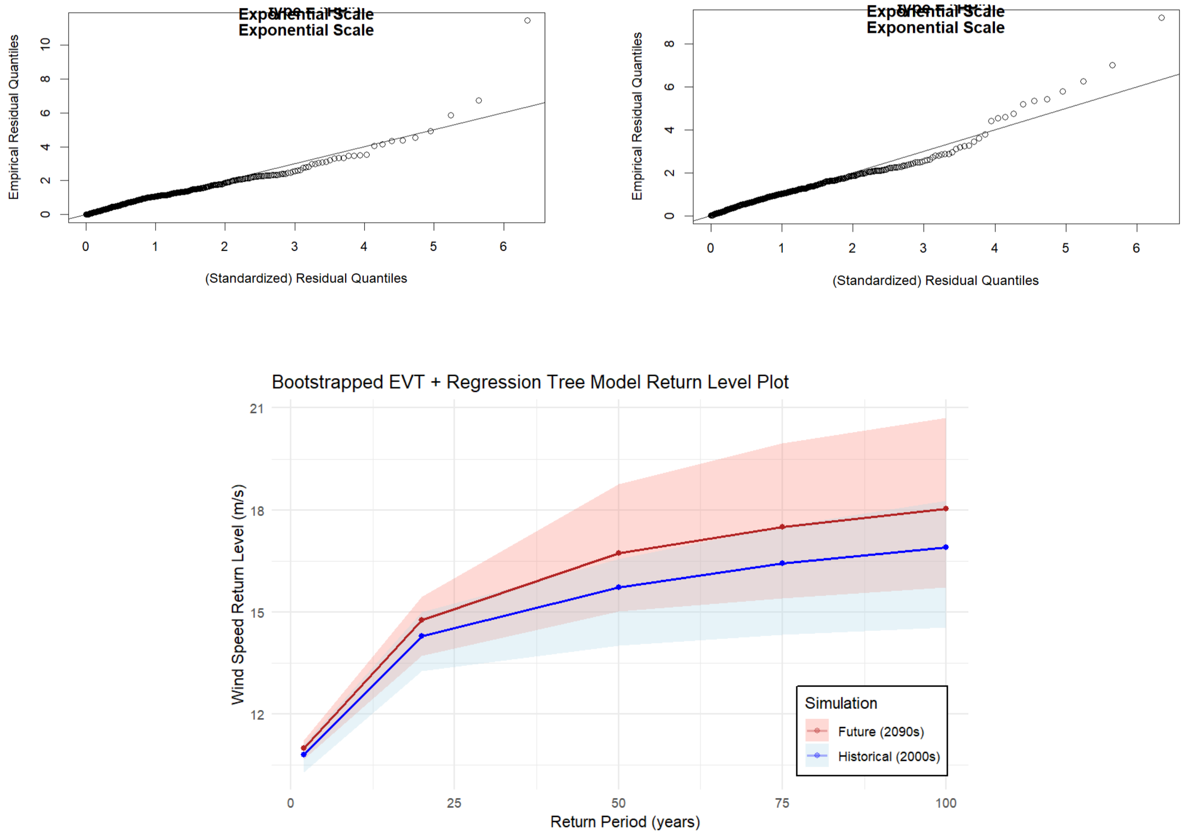


Figure 13: *QQ-plots of 2000s simulation (top left) and 2090s simulation (top right) using clusters defined by the regression tree. Both models define $\sigma(X)$ using regression parameters, by the LRT test. Return level plot of both simulations (bottom) indicate a significant increase in return levels in 2090 compared to 2000s. For the 50-year return level, we see an increase of around 1 m/s, which is around a 5% increase.*

A varying threshold has been selected in both simulations as previously, where we pick the 95th percentile cluster surface winds. The QQ-plots show almost no improvement between the reference and our proposed models, but the significant change between the two models is in the comparison in return levels for both simulations, where our results are now consistent with the results shown in (Morris et al. 2024)[6].

4 Discussion

The important results given by our analysis is twofold: using clusters defined by regression trees we can obtain a better fit, as given in our QQ-plot in Figure 5 (2.) compared with Figure 2c) in our ERA5 dataset. In addition, our regression tree-based methodology helped show an increase in extreme surface wind speeds of around 5% in the 50-year return period, which is consistent with a similar study by (Morris et al. 2024), something in which the reference model was unable to show.

In our analysis of both datasets, we have made many heuristical choices. For example, our cluster thresholds of 95% for our varying threshold were chosen as a heuristic, since the threshold stability and mean residual life plots for each cluster were difficult to interpret. Furthermore, our regression tree has been constructed by filtering out wind speeds above a certain heuristic value. Note that the value the data is filtered on constructs a different regression tree and affects the fit of the model. Thus, a future analysis can focus on finding the best value to filter our data such that our model fits best.

In this study, two significant problems have not been solved, as we only introduced a new methodology along with its results. Firstly, from our ERA5 reanalysis data, the return levels are not realistic given our empirical data. Statistical simulations that can recreate our scenario and calculate joint return levels are needed to validate our results. Secondly, for our VR-CESM model comparisons between both simulations, a more concrete explanation is required to explain how the model constructed using our methodology can show an increase in surface winds between the 2000s and 2090s simulations, whereas the reference model was not able to show an increase.

For our second problem, we believe that the discrepancy was hinted in the analysis of extreme wind speeds in the summer and non-summer months. We showed that there is a significant climate change signal for extreme wind speeds, especially in April and in October, despite a decrease in November between the 2000s and 2090s simulations. It may be possible that we can see more significant discrepancies when we analyze individual clusters between simulations which can explain the increase in return levels.

For this study we have only shown the results of the VR-CESM dataset by only comparing the return levels of wind extremes. We can provide a more thorough analysis by analyzing the statistical behavior of each cluster in each simulation, and look for patterns that may explain the discrepancies between the reference and proposed models. In addition, we can also make an extra assumption on the underlying processes in the climate models, where the clustering by regression trees is the same in both simulations. Although it will not necessarily provide us accurate information on return levels, we can gain better interpretability for return levels, and better analysis can be made for each cluster.

An interesting property of using regression trees in extreme value modeling outside of modeling accuracy is that we can understand better where the uncertainty of surface wind extremes come from. For instance, in our VR-CESM models for each simulation we can see that the atmospheric stability in the first layer helped us differentiate data points that are less extreme to data points that are relatively more extreme. For each dataset we used low level winds (winds at 850hPa), which turns out to be the most significant covariate for extreme surface winds in our regression tree analysis, however it is quite obvious that winds that are slightly above us by a couple of kilometers will be heavily correlated with surface winds on the ground. Such covariates are useful in modeling extreme winds accurately, however, a future research question can be made by analyzing another, less obvious covariate, and determining whether or not there is some significant relationship between the covariate and surface wind.

Lastly, it is worth reminding that we used ANOVA to form our regression tree. We showed that this method performs quite well as a substitute to using the negative log-likelihood as the loss function for our regression tree. We may expect a better fit using the negative log-likelihood due to the nature of this particular loss function.

5 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] Jichao Wang et al. "Spatiotemporal variations and extreme value analysis of significant wave height in the South China Sea based on 71-year long ERA5 wave reanalysis". In: *Applied Ocean Research* 113 (2021).

[2] P.C.D. Milly et al. "Stationarity Is Dead: Whither Water Management?" In: *Science* 319 (2008), pp. 573–574.

[3] Richard L. Smith et al. "Markov chain models for threshold exceedances". In: *Biometrika* 84 (1997), pp. 249–268.

[4] Leo Breiman et al. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[5] Lorenzo Mentaschi et al. "Non-stationary Extreme Value Analysis: a simplified approach for Earth science applications". In: *Hydrology and Earth System Sciences Discussions* 127 (2016), pp. 353–369.

[6] Michael Morris et al. "Resolution-Dependence of Extreme Wind Speed Projections in the Great Lakes Region". In: *Journal of Climate* 37 (2024), pp. 3153–3171.

[7] Michalis I. Vousdoukas et al. "Projections of extreme storm surge levels along Europe". In: *Climate Dynamics* 47 (2016), pp. 3171–3190.

[8] Sebastien Farkas et al. "Generalized pareto regression trees for extreme event analysis". In: *Journal of Extremes* 27 (2024), pp. 437–477.

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

[10] A.C. Davison and R.L. Smith. "Models for Exceedances over High Thresholds". In: *Models for Exceedances over High Thresholds* 52 (1990), pp. 393–442.

[11] Jonathan A. Tawn Emma F. Eastoe. "Modelling Non-Stationary Extremes with Application to Surface Level Ozone". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 58.1 (2009), pp. 25–45.

[12] Christopher A.T. Ferro and Johan Segers. "Inference for Clusters of Extreme Values". In: *Journal of the Royal Statistical Society* 65 (2003), pp. 545–556.

[13] Petra Friederichs. "Statistical downscaling of extreme precipitation events using extreme value theory". In: *Journal of Extremes* 13 (2010), pp. 109–132.

[14] Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Elsevier, 2018.