ONR Internship

Introduction to Extreme Value Theory and Constructing Hazard Curves

███████████

████████████████████

Lancaster University

16th January 2017

1 Introduction

We are often interested in assessing the probability of rare events, from extremely high or low temperatures to excessive rainfall. The rarity of these phenomena means that we may have never before observed events as extreme, and as such, standard statistical techniques cannot be used due to the extrapolation that would be involved. For instance, suppose the highest temperature ever recorded at a particular location is 30°C. Just looking at the data, one may conclude the probability of seeing a temperature higher than this to be 0, but this is unlikely to be true. To tackle problems such as this, techniques from Extreme Value Theory can be used. These methods provide rigorous methods to model rare events. We begin by introducing the methods used to model individual variables, before discussing the notion of multivariate extremes and some of the techniques that can be used to model multiple variables, or the joint probabilities of rare events.

2 Univariate Extreme Value Theory

To begin, we will only consider modelling extreme values that are large, as the methods used to model small values can be obtained from the same theory. There are two main ways to model the extreme values of a single variable. The first is to split a data set into sections of equal length (corresponding to, for example, one week or one year's worth of observations) and take the maximum value observed in each section, as demonstrated in Figure 1 for some example data of daily temperature readings across ten years at a single location. This collection of maximum values is known as a set of block maxima for the data set, and these can be modelled using what is known as the Generalised Extreme Value distribution.
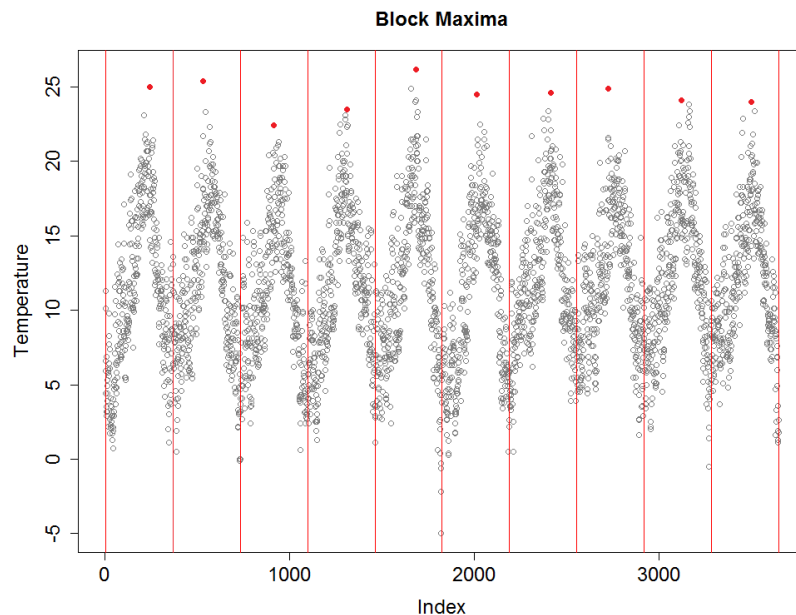

Figure 1: Example of the block maxima approach to defining extremes for one variable.

The second method is to consider all values above a given threshold. In this case, it is the Generalised Pareto distribution that can be used to model the data. This method is often preferred over the block maxima approach because more values that are considered extreme are

taken into account. For instance, in the block maxima approach with blocks of one year, there could be a particularly large observation in both June and September, but only one of them would be considered in the model, whereas the threshold exceedance method takes both of these events into account. This method is demonstrated in Figure 2, again using temperature data across ten years at a single location.
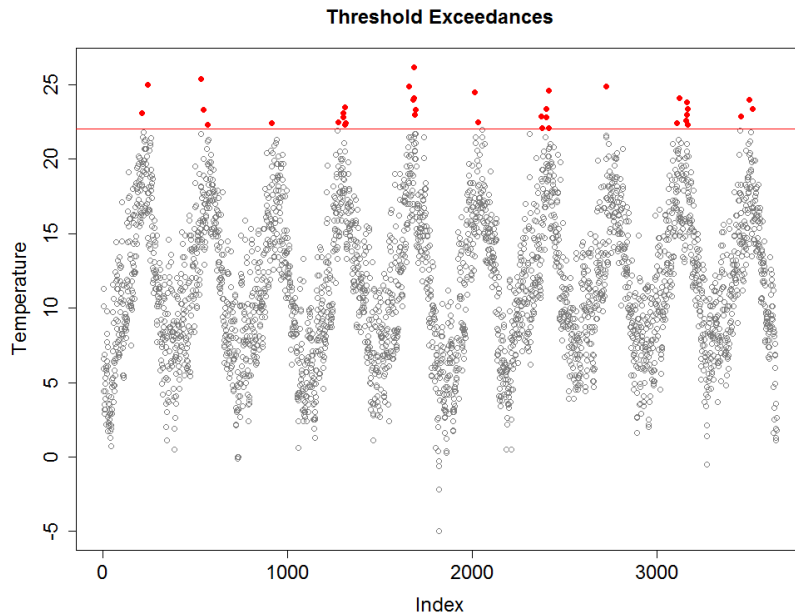

Figure 2: Example of the threshold exceedance approach to defining extremes for one variable.

In the methods we will be using, we are only interested in the threshold exceedance technique, which we will now discuss in more detail. Suppose that we have chosen to model the variable X above the threshold u, then the Generalised Pareto distribution has the form

$$\Pr(X > x | X > u) = \left[1 + \frac{\xi(x - u)}{\sigma}\right]^{-1/\xi},$$

for $x > u$, and some values of the parameters ξ and $\sigma > 0$. We can estimate the parameters of this model for a given set of data using packages already available in R. There is a large amount of theory around choosing an appropriate threshold, u when using the GP distribution, but it is common practice to use a specified high quantile of the data. For instance, we could choose u to be the 95th quantile of X, i.e. the value $x_{0.95}$ that satisfies

$$\Pr(X < x_{0.95}) = 0.95,$$

which may be found empirically from the data. If we set the threshold to be $u = x_{0.95}$, then the probability that X is greater than some value $x > u$ can be calculated using

$$\Pr(X > x) = \Pr(X > x | X > u) \cdot \Pr(X > u)$$
$$= \Pr(X > x | X > u) \cdot 0.05,$$

which can then be calculated completely once the parameters (ξ, σ) of the model have been estimated. Note that a quantile other than the 95th can be used, and threshold choice should be

2

supported by diagnostic plots.

For more information on univariate extreme value techniques, the book by Coles (2001) is a useful resource.

3   Multivariate Extreme Value Theory

When we begin to consider more than one variable, there are several ways to define an extreme event. We begin by considering the bivariate case. Given two variables X and Y, one option is to consider the probability that X is above some value x at the same that Y exceeds some value y, i.e. $\Pr(X > x, Y > y)$, where at least one of x or y is large. Using standard probability theory, it is possible to separate this into two components, one being a univariate probability as discussed in the previous section, and the other a conditional probability. This would be written as

$$\Pr(X > x, Y > y) = \Pr(Y > y | X > x)\Pr(X > x).$$

In practice, if x is large, the univariate component, $\Pr(X > x)$, can be estimated using the Generalised Pareto distribution that was discussed in the previous section. For the other component, the probability that Y is greater than y given that X is greater than x, we can use the methods developed by Heffernan and Tawn (2004).

As with most methods in multivariate extreme value theory, Heffernan and Tawn's approach requires transformation of the variables X and Y to some common, standard distribution. To do this, we can use what is known as the Probability Integral Transform, for which we require estimates of the distribution functions $F_X(x) = \Pr(X \leq x)$ and $F_Y(y) = \Pr(Y \leq y)$ of the variables X and Y respectively. To estimate the distribution functions, it is common to use the empirical distribution of each variable below some threshold, and to make use of the Generalised Pareto distribution above that threshold to estimate the tail behaviour of each random variable. When applying Heffernan and Tawn's method, it is a common choice to use Laplace margins, which can be achieved using the transformation

$$X_L = \begin{cases} \log\{2F_X(X)\} & , \ F_X(X) \leq \frac{1}{2}, \\ -\log\{2(1 - F_X(X))\} & , \ F_X(X) > \frac{1}{2}, \end{cases} \tag{1}$$

for the random variable X. We obtain the transformation Y_L of the random variable Y in an analogous way.

Figure 3 demonstrates how data may look once it has been transformed to Laplace margins using the method described above. Here, we use the example of temperature and wind speed data which will be discussed further later in this document.

Once this transformation has been applied, Heffernan and Tawn show that for a wide class of distributions, the variables

$$\frac{Y_L - \alpha X_L}{X_L^\beta} \quad \text{and} \quad X_L - u$$

are approximately independent for $X_L > u$, and that $X_L - u | X_L > u$ follows an exponential distribution (that is, a GPD with parameters $\sigma = 1$ and $\xi = 0$).
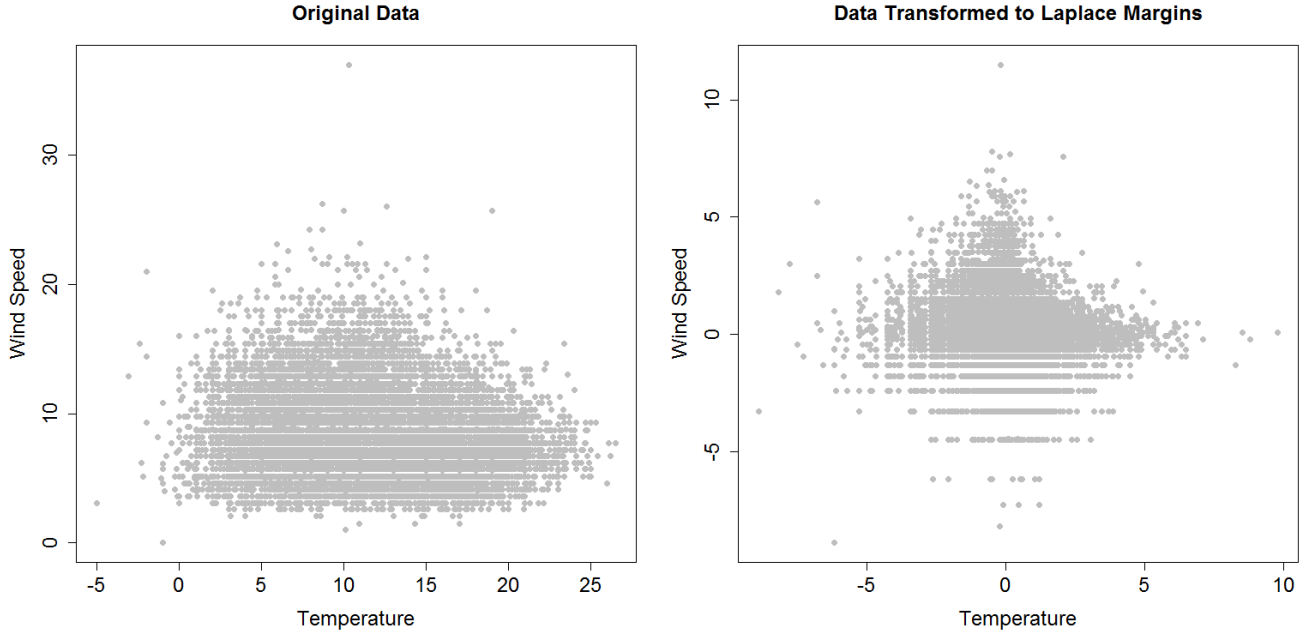
Figure 3: Example of transforming data to Laplace margins.

The parameters α and β can be estimated using non-linear regression techniques. In particular, we may consider a model of the form

$$Y_L = \alpha X_L + X_L^{\beta} Z,$$

where Z represents the residuals of the model. In this case, we often make the working assumption that Z follows a Gaussian distribution, i.e. $Z \sim N(\mu, s^2)$. The estimates obtained from the non-linear regression are usually denoted by $(\hat{\alpha}, \hat{\beta})$.

With estimates $\hat{\alpha}$ and $\hat{\beta}$, we can simulate data that are more extreme than our sample, which allows us to estimate probabilities of extreme events. To simulate new values $(x_{L,new,1}, y_{L,new,1})$, ... , $(x_{L,new,m}, y_{L,new,m})$ for X above a threshold u, the algorithm is:

1. set $x_{L,new,i} = u + e_i$, where $e_i \sim \text{Exp}(1)$;

2. sample z_i from the empirical distribution of the observed residuals (with replacement);

3. set $y_{L,new,i} = z_i (x_{L,new,i})^{\hat{\beta}} + \hat{\alpha} x_{L,new,i}$;

for $i = 1, ..., m$. The resulting simulated values can be transformed back from the Laplace margins to their original scales using the inverse of Equation (1).

Figure 4 gives an example of simulated data that can be obtained using the method outlined above. Again, this example uses temperature and wind speed data. The original data is shown by the grey points, the red line shows the threshold above which we are simulating data (chosen to be the 99.5th quantile of the temperature values in this case), and the green points show the simulated temperature and wind speed values.
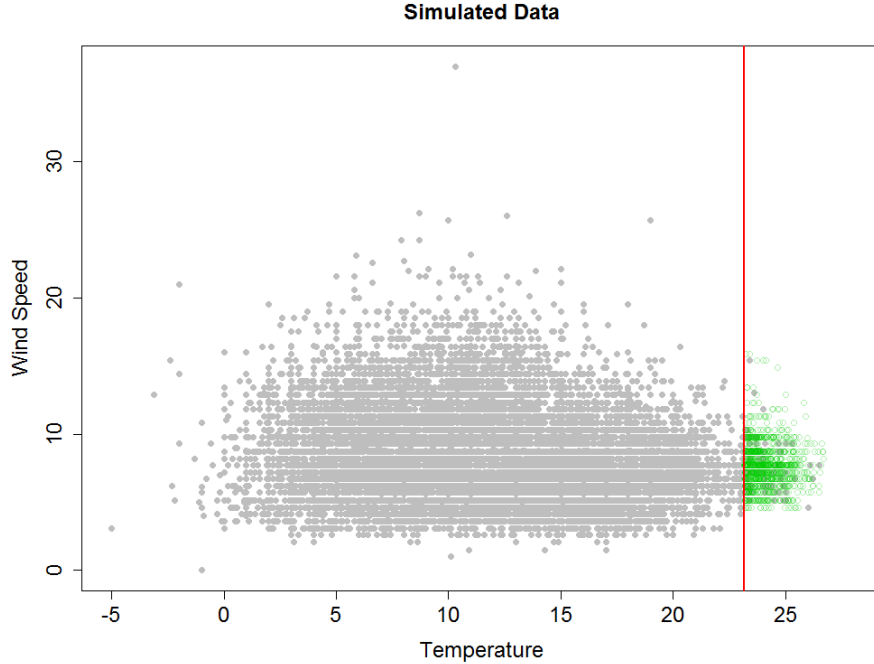
4

Figure 4: Example of simulating data given that one variable is large using Heffernan and Tawn's method.

This approach to modelling multivariate extremes is particularly useful as it can be applied in a variety of scenarios. While some other methods for modelling multivariate extremes can only be used if there is either asymptotic dependence or asymptotic independence between the variables, Heffernan and Tawn's method can be used in both instances. That is, it doesn't matter whether extreme values occur simultaneously in both variables or not. Coles et al. (1999) and Chapter 9 of Beirlant et al. (2004) provide further discussion of asymptotic dependence and asymptotic independence. It is also usually possible to use Heffernan and Tawn's approach when there is negative dependence between the variables, which is not always the case with other methods.

Heffernan and Tawn's method can be applied to a data set using the `texmex` (2013) package in R. Functions within this package allow us to fit Heffernan and Tawn's model and produce estimates of $\Pr(X > x, Y > y)$. For a given value of x, we are able to produce simulated data for the two variables given that X is greater than x. Using this simulated data, it is possible to find an empirical estimate of $\Pr(Y > y | X > x)$. Finally, multiplying this by an estimate for $\Pr(X > x)$ obtained using the Generalised Pareto distribution gives an estimate of $\Pr(X > x, Y > y)$.

It is important to note that either variable can be used as the conditioning variable in this method as long as it is large. In other words, the probability can be alternatively written as

$$\Pr(X > x, Y > y) = \Pr(X > x | Y > y)\Pr(Y > y).$$

There may be situations where it is preferable to use one variable over the other as the conditioning variable, as will be the case in the work we will discuss in later sections.

A further advantage of Heffernan and Tawn's method over other techniques is that it is reasonably straightforward to extend the methods to model more than two variables. Suppose we are interested in modelling three variables, X, Y and Z. Heffernan and Tawn's method allows simulation of a new trivariate data set conditioning on any one of the variables being extreme. In particular, if we consider the formula

$$\Pr(X > x, Y > y, Z > z) = \Pr(Y > y, Z > z | X > x)\Pr(X > x),$$

then we can estimate $\Pr(X > x)$ for x large using the Generalised Pareto distribution, as before, and simulate values using Heffernan and Tawn's method to obtain an estimate of $\Pr(Y > y, Z > z | X > x)$. The approach also extends in a similar manner for higher than three dimensions.

As an alternative for considering $\Pr(X > x, Y > y, Z > z)$, it may also be possible to subset the data and consider only values of X and Y corresponding to $Z > z$ for some specified value of z (which may be moderately extreme). This is suitable in situations where we have large amounts of data, as we require there to be sufficient data left to carry out the methods once only those points corresponding to $Z > z$ are considered. In this case, the problem becomes bivariate again, as and the theory can be applied as before.

4 Constructing a Bivariate Hazard Curve

One of the aims of this project is to develop a method to produce plots showing combinations of values of our variables that are expected to be exceeded once every 100 or 10,000 years. Suppose we have daily observations of the random variables X and Y, as will be the case in the data investigated in this project. This means that we want to find values of (x, y) such that

$$\Pr(X > x, Y > y) = p,$$

where $p = \frac{1}{365} \cdot 10^{-2}$ (corresponding to a 1 in 100 year event) or $p = \frac{1}{365} \cdot 10^{-4}$ (corresponding to a 1 in 10,000 year event), assuming that there is independence between the observations on different days. Once we have found pairs of values (x, y) satisfying this equation, we can plot them on a graph, and construct what we call a 'bivariate hazard curve' for the data. We can use the models described in the previous sections to find such points.

Suppose that variable X is temperature and variable Y is wind speed. First, we fit a Generalised Pareto distribution to the temperature variable, and find the qth quantile of this distribution for a range of values of q close to 1. These quantiles provide the x values of our points. Heffernan and Tawn's model is then fitted to the data and simulations of temperature and wind speed data are obtained from this given that temperature is above $x°$C. Using these simulations, we can estimate a wind speed value y that satisfies our requirement that $\Pr(X > x, Y > y) = p$, for the value of p we are interested in. For some of the points used to form the hazard curve, this is done the other way around, i.e. we fit a GP distribution to the wind speed variable, and the qth quantile of this distribution becomes the y value of one of the points used to create the curve. We then use Heffernan and Tawns approach to simulate wind speed and temperature data given that wind speed is above y m/s, and find the corresponding x value that gives us $\Pr(X > x, Y > y) = p$.

In the univariate case, when constructing a hazard curve (or return level plot) for the variable X, we would be interested in pairs of values (p, x_p) satisfying

$$\Pr(X > x_p) = p.$$

The hazard curve would join several points satisfying this for many different values of p. Here, when we construct a bivariate version, we are only considering curves corresponding to individual values of the probability p. As such, it should be noted that these bivariate hazard curves really correspond to 'slice' of a surface that could be obtained for values (p, x_p, y_p) satisfying

$$\Pr(X > x_p, Y > y_p) = p,$$

where instead of considering multiple values of p simultaneously, we simply consider one value at a time.

4.1   Bivariate Hazard Curve Examples

In this subsection, we discuss an example bivariate hazard curve, and explain how it should be interpreted. The data used in this document is from the website of the Centre for Environmental Data Analysis (`http://catalogue.ceda.ac.uk/uuid`). The data set consists of historical readings of a variety of weather variables at various weather stations across the world. In this illustrative example, we consider temperature and wind speed data for one weather station on the north east coast of England (latitude 55.42, longitude -1.6) between 1973 and 2011. The data for the temperature and wind speed variables is given as hourly readings. For ease of computation, daily maximum values have been used to compute the bivariate hazard curves discussed here. The data contained some missing values, which have been ignored.
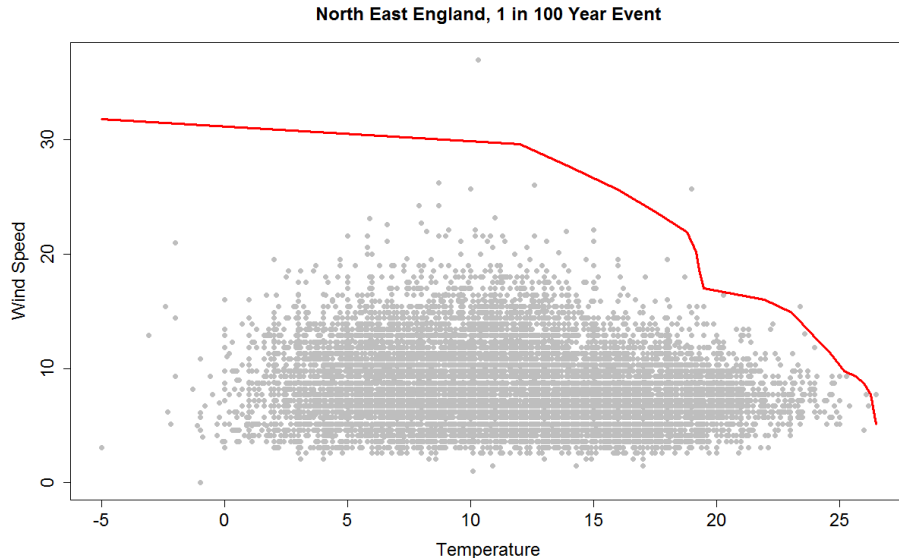

Figure 5: Plot of temperature and wind speed data for location in the north east of England, with the hazard curve for a 1 in 100 year event shown in red.

Our bivariate hazard curve is given by the red line in the plot in Figure 5. It shows combinations of temperature and wind speed values that we estimate to be exceeded once every 100

years (equivalent to a yearly probability of exceedance of 10^{-2}).

If a point (x, y) lies on the hazard curve, this means that the probability of seeing a maximum daily temperature reading of at least $x°$C with a maximum daily wind speed reading of at least y metres per second in any given year is 10^{-2}.

Examples of this are given in the plots in Figures 6. In the left hand plot of Figure 6, the point (22, 16) is shown to lie on the hazard curve, which means that the probability of seeing one day in a given year where the maximum temperature is greater than 22°C and the maximum wind speed is greater than 16m/s is 10^{-2}. This is equivalent to saying that we would expect to see a temperature and wind speed event that lies within the shaded region of the plot once every 100 years.

Similarly, in the right hand plot of Figure 6, the point corresponding to a temperature of 15°C and a wind speed of 26.6m/s is shown to lie on the hazard curve. Again, the probability that we simultaneously exceed these values (i.e. that an event lies in the shaded region) once in a given year is 10^{-2}.
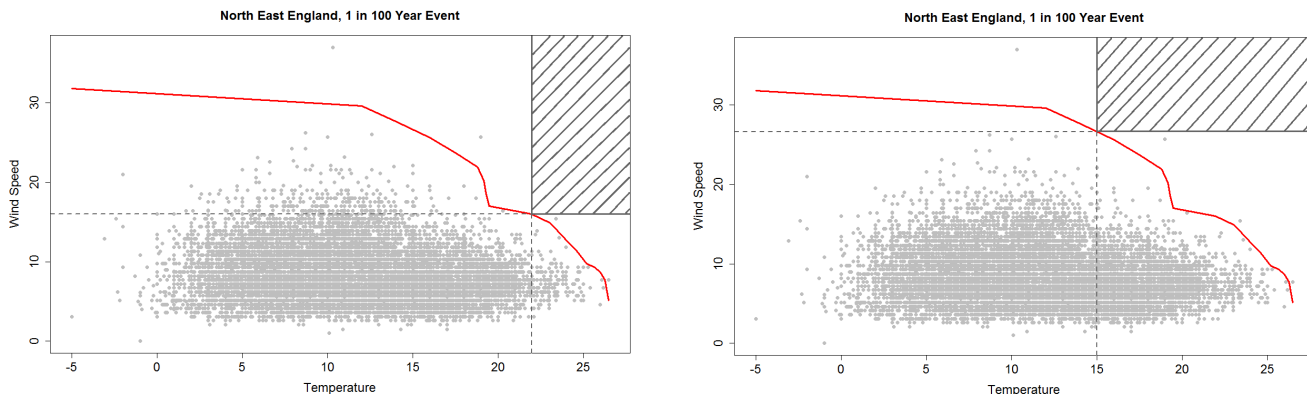

Figure 6: Examples of reading the hazard curve for the temperature and wind speed data.

It is important that the hazard curve is interpreted correctly, as explained above. In particular, it should be noted that it is not the case that there is a probability of 10^{-2} of lying anywhere in the region to the north east of the curve.

It is also possible to obtain these hazard curves for other probabilities of exceedance. For example, if we were interested in an event that we expect to occur once in every 10,000 years, equivalent to a yearly probability of exceedance of 10^{-4}, we would obtain the plot in Figure 7, which should be interpreted in a similar manner to the previous examples.

4.2 Uncertainty in Hazard Curves

There is no standard way to represent uncertainty in bivariate plots, but a method often used to quantify uncertainty more generally in statistics is bootstrapping. This is a method we have
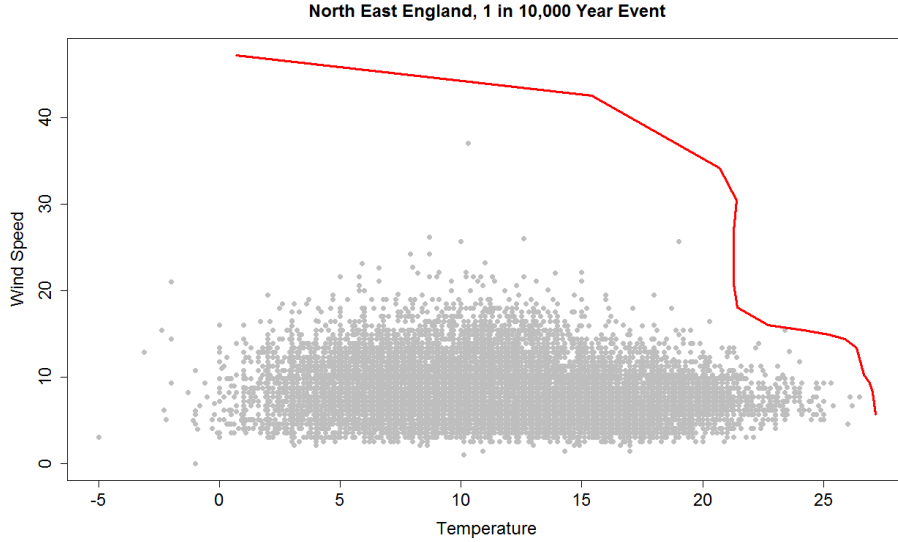
Figure 7: Plot of temperature and wind speed data for location in the north east of England, with the hazard curve for a 1 in 10,000 year event shown in red.

applied to our bivariate hazard curve calculations, in order to give some idea of the uncertainty in the results.

To bootstrap, we use the following algorithm:

1. Take a random sample of points from the original data set (with replacement), of the same size as the original data set.

2. For this new set of sampled data, calculate the same statistic that we were interested in from the original data (e.g. the mean of the data).

3. Repeat steps 1 and 2 a large number of times (usually 1000 in the examples in this document).

On completion of the bootstrapping procedure, we will have obtained a number of bootstrapped values for the statistic of interest, and these bootstrapped values give us an idea of the sampling variability of the statistic. In particular, the bootstrapped values can be used to obtain confidence intervals for the statistic. For example, a 95% confidence interval for the statistic is provided by the central 95% of bootstrapped values.

Let's once again consider our bivariate hazard curve for the temperature and wind speed data with an annual exceedance probability of 10^{-2}. As discussed in the previous section, this curve is constructed by finding several points (x, y) satisfying $\Pr(X > x, Y > y) = \frac{1}{365} \cdot 10^{-2}$, where X represents maximum daily temperature and Y is the maximum daily wind speed variable. Once several of these points have been found, we can join them together to form our hazard curve. We can use bootstrapping to quantify uncertainty in the points that we use to construct the curve.

When constructing a hazard curve, there is uncertainty in exactly which pairs of values (x, y) satisfy $\Pr(X > x, Y > y) = p$, and this is where we can use bootstrapping techniques. As

discussed above, in bootstrapping, we take samples of data (with replacement) from our original data set, and calculate the same statistics for this sampled data as we did for the original data. So in the case of the hazard curve, once we have a bootstrap sample, we follow the same procedure as in Section 4 to find wind speed and temperature values that have a yearly joint probability of exceedance of 10^{-2}.
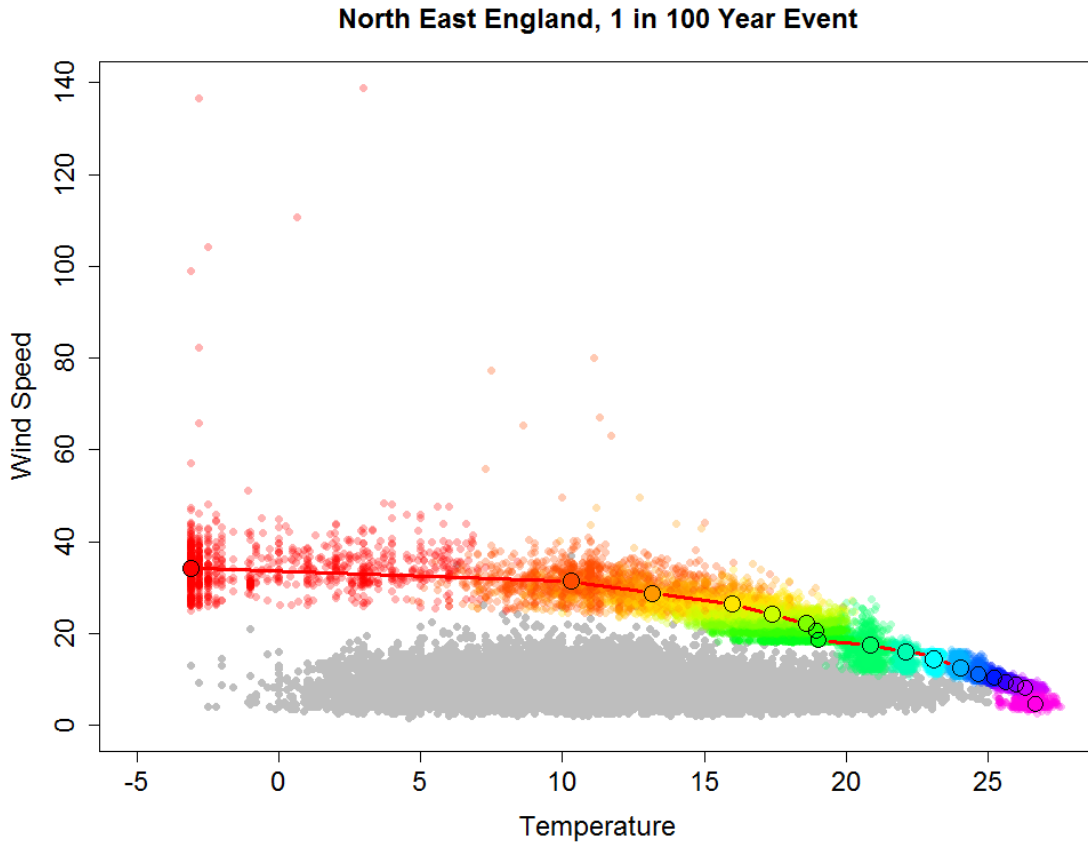

Figure 8: Plot of temperature and wind speed data for location in the north east of England, with the hazard curve for a 1 in 100 year event shown in red and bootstrapped values shown by the coloured points.

We carry out this sampling and calculation method 1000 times to give us an idea of the uncertainty present. The plot in Figure 8 demonstrates this method. The solid red line shows our original hazard curve, and the large circular points on this line shows the points obtained from the original data that have been used to construct this hazard curve. Each of these points has a different colour, and the smaller points with the same colour show the values obtained by bootstrapping 1000 times conditioning on the same variable and using the same quantile q as were used to obtain the original points on the curve.

These bootstrapped points suggest that there is more variability in the points with low temperature and high wind speed than with high temperature and low wind speed, which may be due to the difference in the heaviness of the tails of the marginal distributions of the temperature and wind speed variables.

As mentioned previously, we can use our bootstrapped points to form a sort of confidence region for our hazard curve. This is demonstrated in Figure 9. In this case, we consider just one of the points used to construct the hazard curve, which is given by the red point in the plot. The grey points show bootstrapped values corresponding to this point. To construct what we will call a 95% confidence plot, we find 95% confidence intervals separately from the x and y bootstrapped values, by taking the central 95% of values of each according to the quantiles. These confidence intervals are shown by the green lines on the plot, and form a rectangular shape. We are interested in the top right corner and bottom left corner of this rectangle, as shown by the blue points. The top right corner corresponds to an upper confidence point and the bottom left to a lower confidence point. Once these points have been found for all values used to construct the hazard curve, all those corresponding to upper confidence points can be joined together, and the same can be done for the lower confidence points. These lines together form our confidence region.
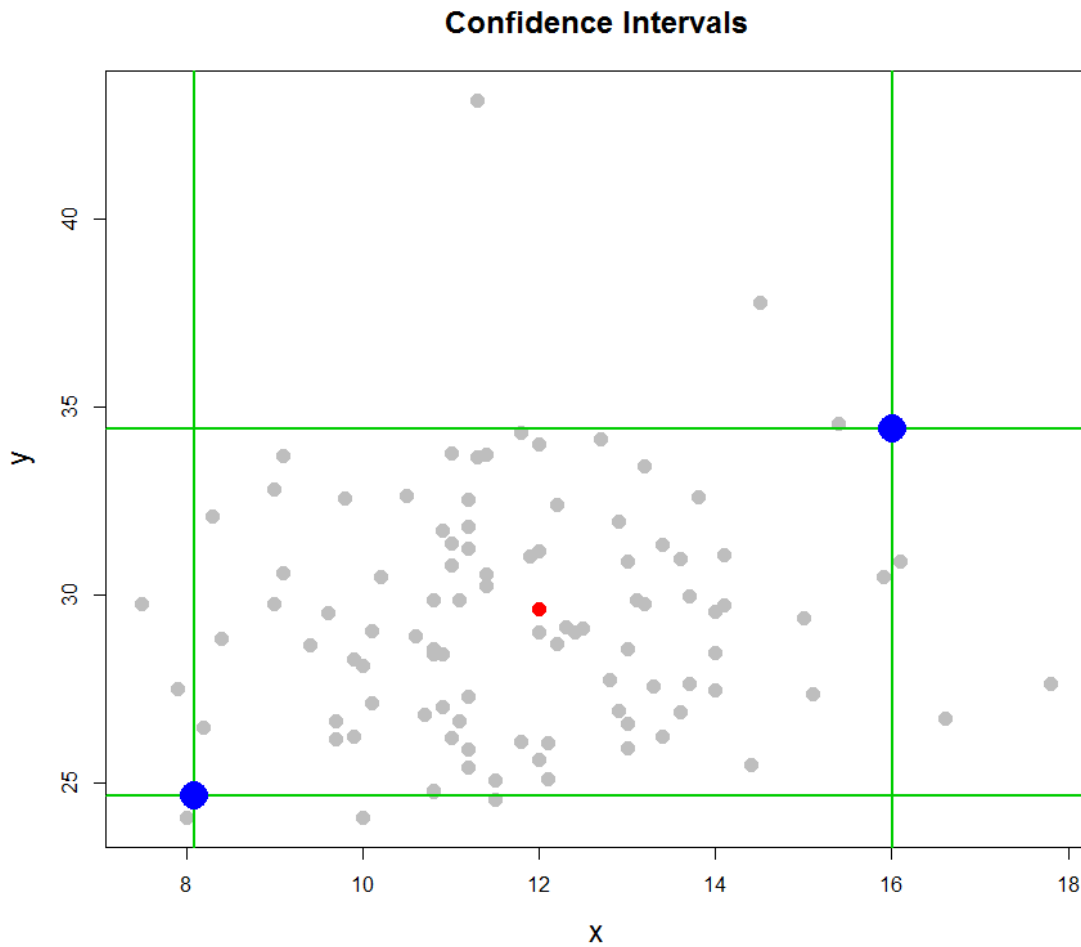

Figure 9: Example of using bootstrapped values to construct confidence intervals.

The plots in Figure 10 show 68.3% (corresponding to one standard deviation either side of the mean in a Normal distribution), 95% and 99% confidence regions for the temperature and wind speed hazard curve with a hazard curve showing a yearly probability of exceedance of 10^{-2}.
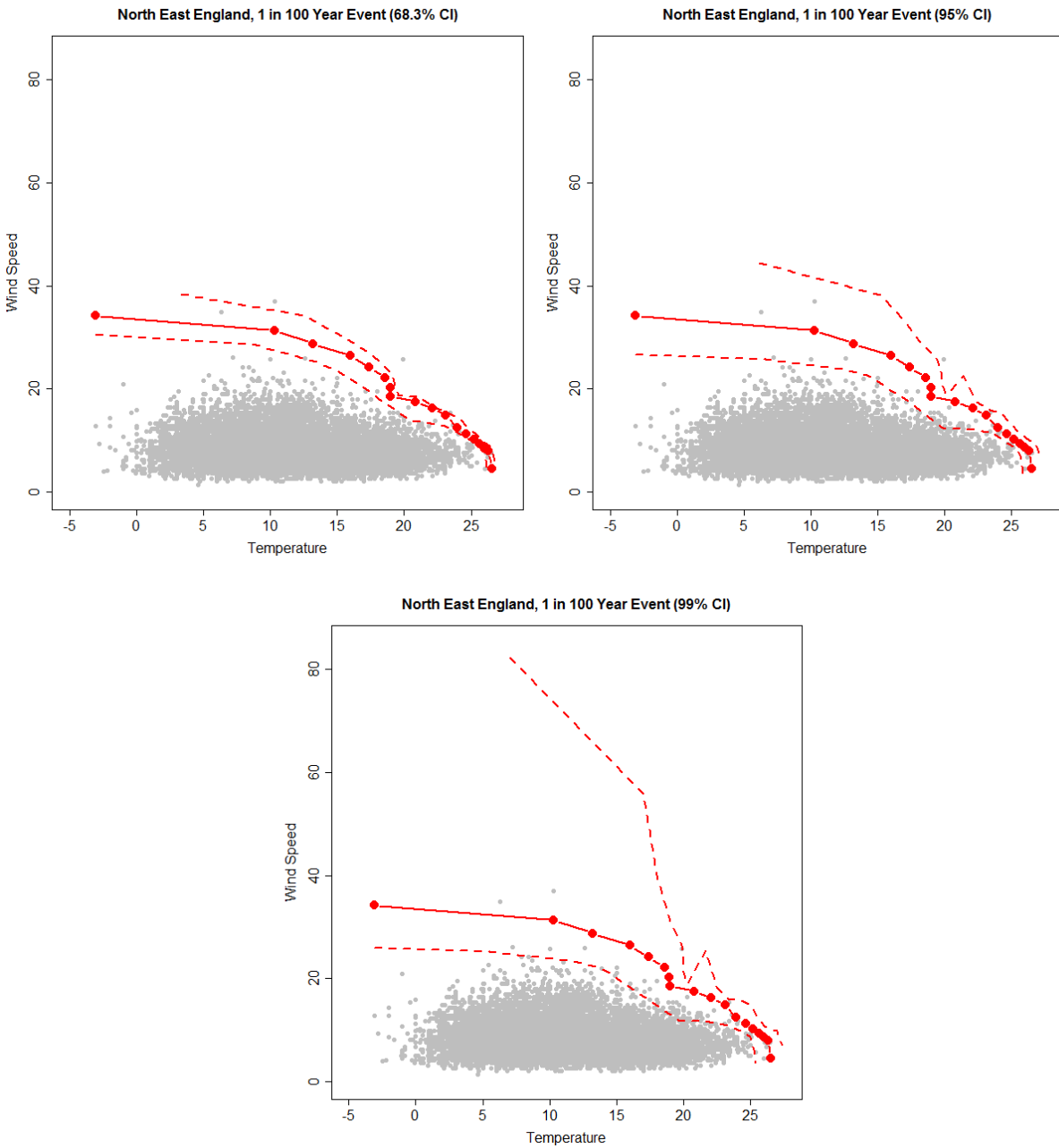

Figure 10: Bivariate hazard curves for temperature and wind speed data. Three different confidence regions are given by the red dotted lines.

As mentioned previously, for each point used to create the original hazard curve, we have 1000 bootstrapped points. To find the corresponding confidence regions, we find confidence intervals for the X and Y variables separately and plot the corresponding upper confidence bounds and lower confidence bounds on our plot together. We repeat this for each of the points used to obtain the original curve, and finally join all the upper confidence bounds and all the lower confidence bounds obtained to give our upper and lower confidence bounds on the hazard curve.
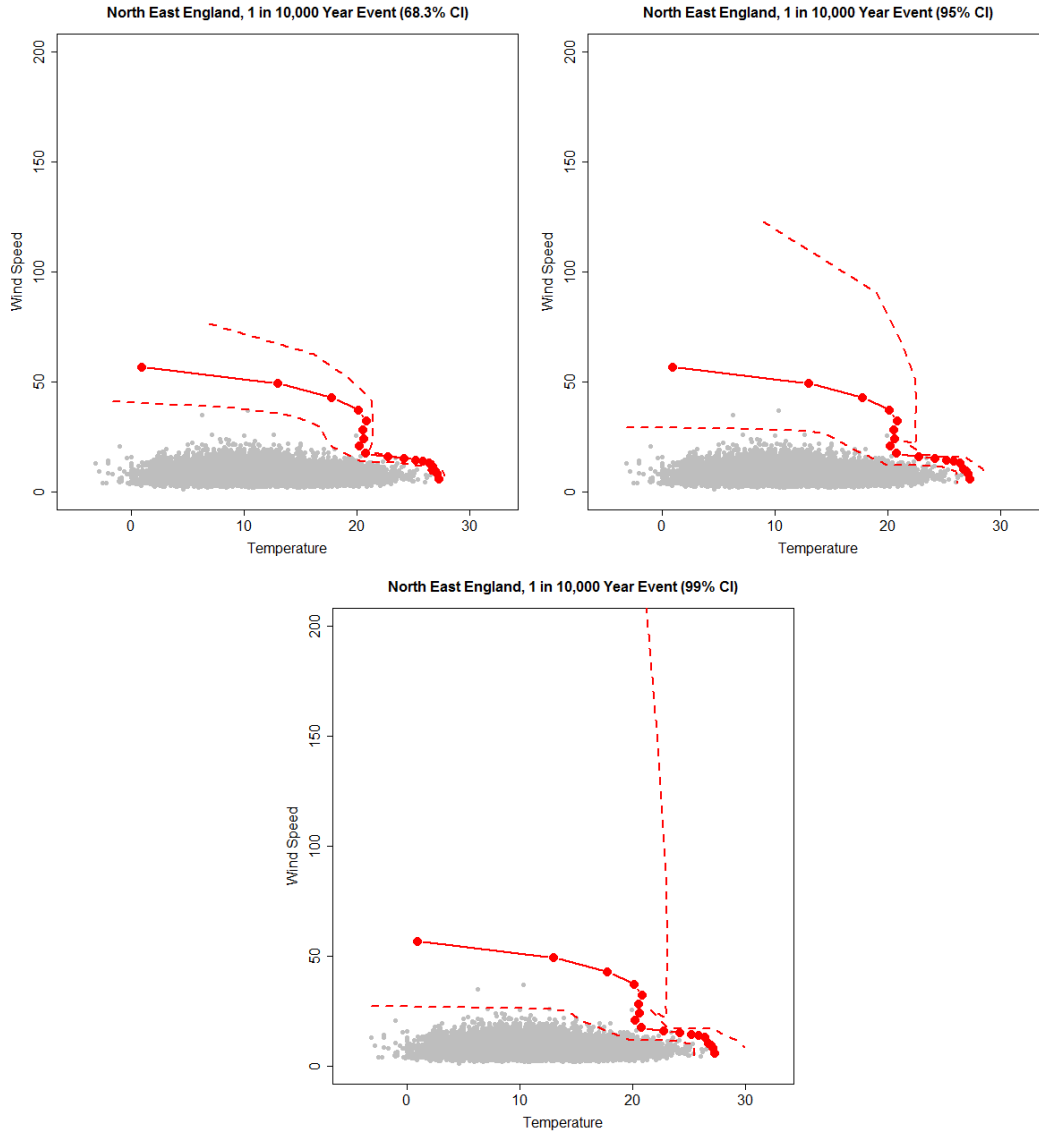

Figure 11: Bivariate hazard curves for temperature and wind speed data for values expected to be exceeded once every 10,000 years. Three different confidence regions are given by the red dotted lines.

We can also obtain these confidence regions for the 10^{-4} hazard curve that was discussed earlier, as shown in Figure 11. Here, the confidence regions are much wider than for the 10^{-2} hazard curve because more extrapolation has to be done to estimate 1 in 10,000 year events than 1 in 100 year events, which increases the uncertainty in the model.

4.3   Hazard Curves for Low Temperature and High Wind Speed

Also of interest are hazard curves for low temperatures and high wind speeds. These can be constructed using the same method as described for high temperature and high wind speed, and should be interpreted in a similar way. To model low temperatures instead of high temperatures, we can transform the temperature variable by taking the negative of each of the observations. Once this transformation is complete, a hazard curve can be constructed by the same method that was used for modelling the high values. The hazard curve that has been obtained can then be transformed back by taking the negative of the transformed values corresponding to temperature. This results in a hazard curve for low temperature and high wind speed. This process is demonstrated in Figure 12.
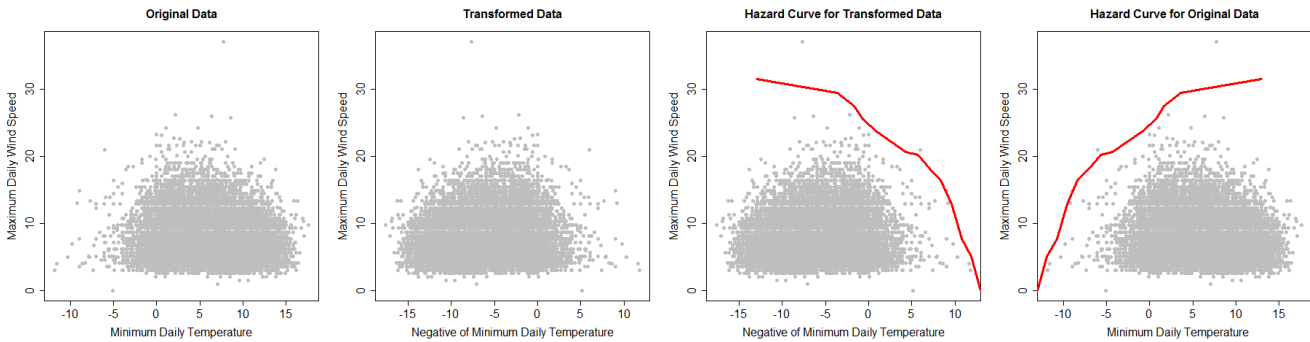

Figure 12: The method for constructing a hazard curve for low temperature and high wind speed.

The left hand plot of Figure 13 shows a hazard curve for one in 100 year events. In this case, if a point (x, y) lies on the curve, it means there is a probability of 10^{-2} of observing a minimum temperature less than $x°$C and maximum wind speed greater than y m/s in any given year. This is further demonstrated by the right hand plot of this figure, in which the point (-5,22.3) is shown to lie on the curve, suggesting that there is a 10^{-2} probability of observing a point in the shaded region, i.e. with a temperature reading below -5°C and a wind speed above 22.3 m/s.
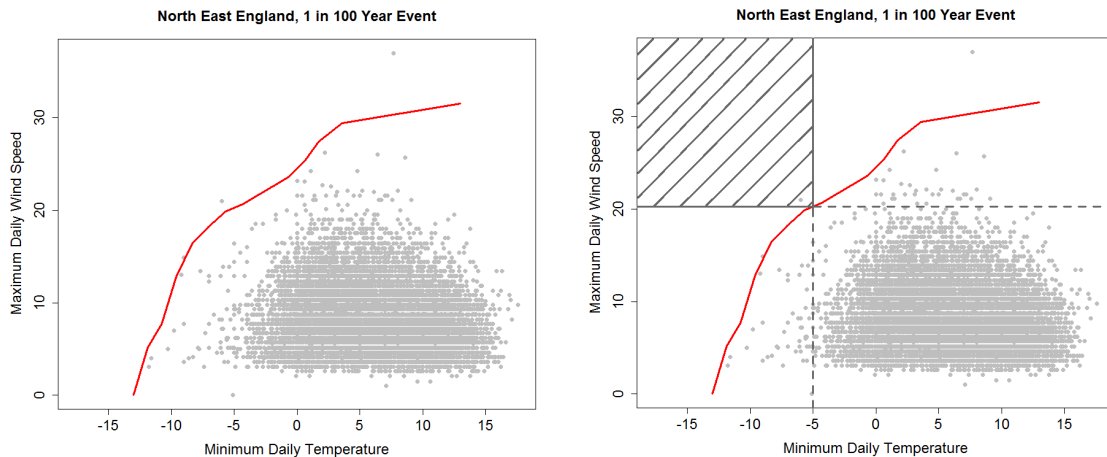

Figure 13: Plot of minimum temperature and maximum wind speed data, with the hazard curve for a 1 in 100 year event shown in red, and an illustration of how this should be interpreted.

As with the previous examples, we can also construct confidence intervals for this hazard curve using bootstrapping techniques. The hazard curve is plotted again in Figure 14, this time with three different confidence regions, corresponding to 68.3%, 95% and 99% confidence levels. To construct these confidence intervals, 1000 bootstrapped values were obtained for each of the points used to construct the initial hazard curve.
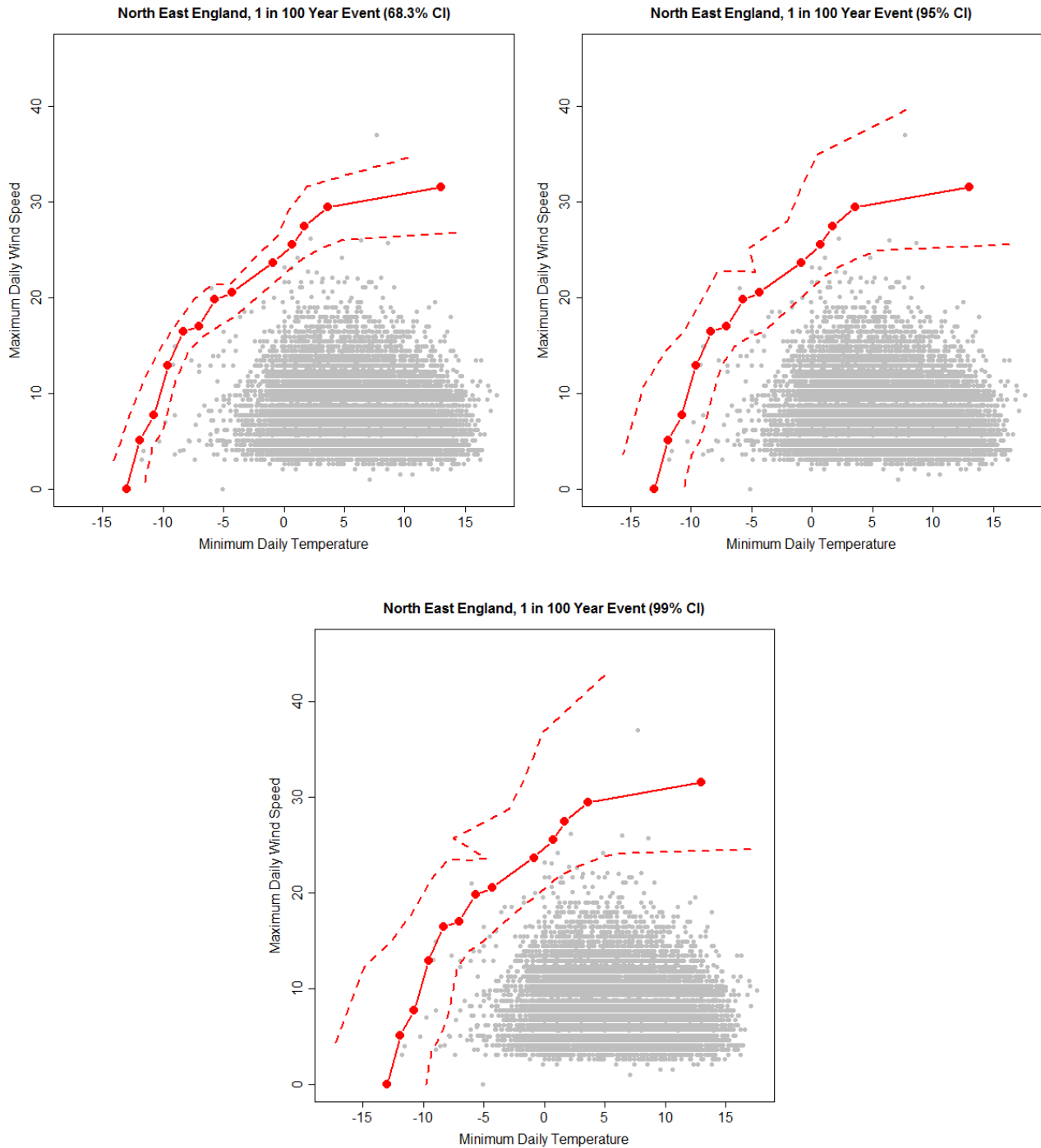

Figure 14: Confidence regions for the minimum temperature and maximum wind speed hazard curve for 1 in 100 year events.

We can also use the same method to construct hazard curves for 1 in 10,000 year events, corresponding to a yearly probability of occurrence of 10^{-4}. Figure 15 demonstrates the hazard curve obtained for 1 in 10,000 year events, with 68.3%, 95% and 99% confidence intervals shown by the dotted lines. Again, these confidence intervals were constructed from 1000 boostrapped values.
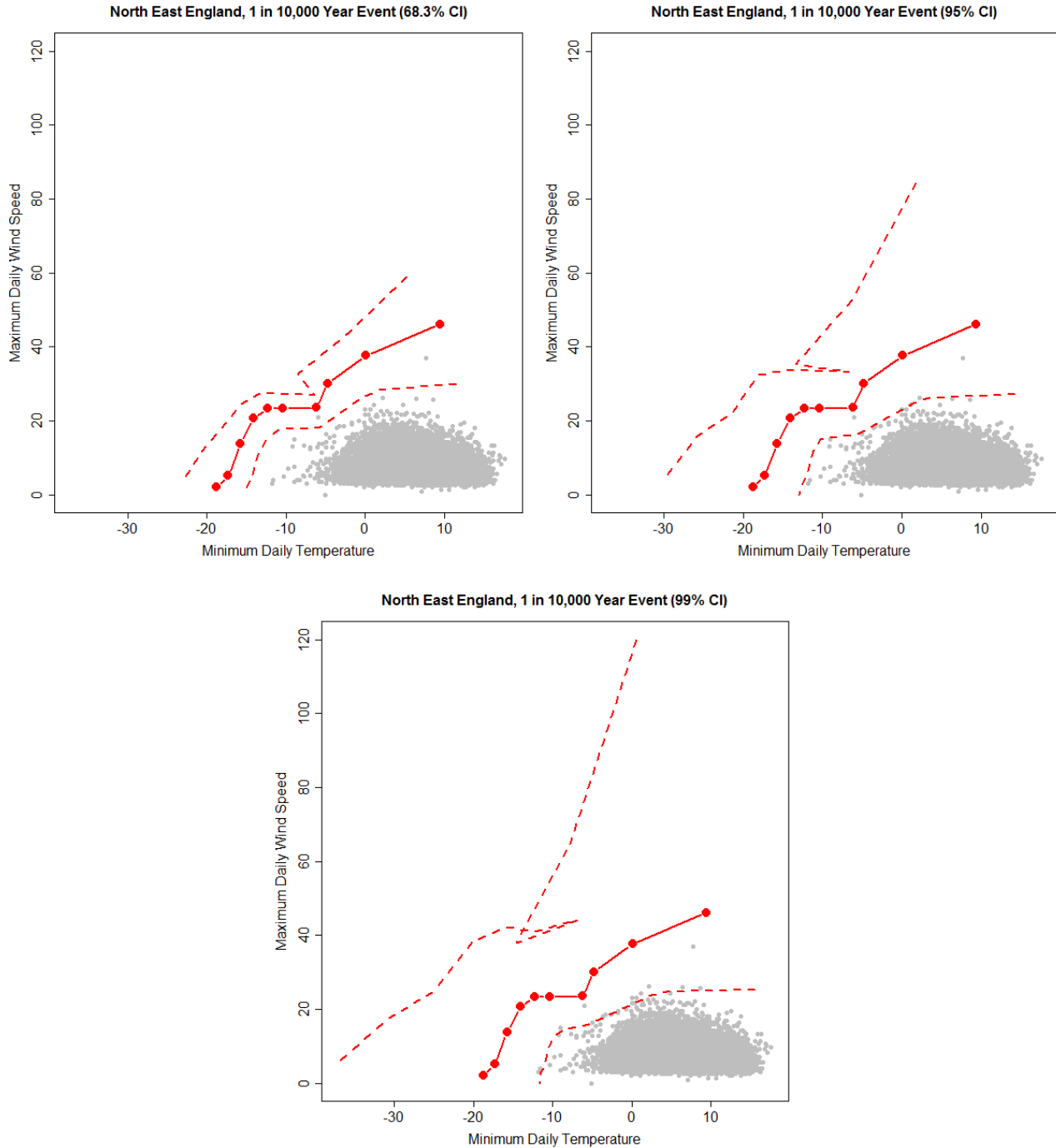

Figure 15: Confidence regions for the minimum temperature and maximum wind speed hazard curve for 1 in 10,000 year events.

4.4 Hazard Curves with Conditioning on One Variable being Extreme

In some cases, we may be interested in more than just the two variables X and Y. In particular, we may wish to introduce a third variable, which we denote Z, and of interest may be finding values (x, y, z) satisfying

$$\Pr(X > x, Y > y, Z > z) = p,$$

for some specified value of p.

As mentioned in Section 3, one option when considering three variables is to only consider a subset of the data corresponding to one of the variables being above some specified value, and use bivariate techniques on this subset of the data.

Suppose we condition on the variable Z being greater than some value z, then we may consider the left hand side of the equation above as

$$\Pr(X > x, Y > y, Z > z) = \Pr(X > x, Y > y | Z > z)\Pr(Z > z),$$

and if we choose z to be some quantile of the variable Z so that $\Pr(Z > z) = r$, say, we have

$$\Pr(X > x, Y > y, Z > z) = \Pr(X > x, Y > y | Z > z) \cdot r.$$

Using this information, finding values of (x,y) for a given z, such that $\Pr(X > x, Y > y, Z > z) = p$ is equivalent to finding values of (x, y) such that

$$\Pr(X > x, Y > y | Z > z) = \frac{p}{r},$$

which is of the same form as our initial problem, but with the data used to calculate the hazard curve being the subset of the original data that corresponds to Z being greater than z.

So far, we have just considered temperature and wind speed in our examples, but another variable that may be of interest is rainfall. We demonstrate the above method using this data. The data consists of daily readings of maximum temperature, maximum wind speed and total rainfall, and in this example we choose rainfall to be our conditioning variable. The plots in Figure 16 show hazard curves for 1 in 100 year events (corresponding to $p = \frac{1}{365} \cdot 10^{-2}$), and we have chosen to condition on the fact that daily rainfall is above its 75th quantile (corresponding to $r = 0.25$). The data plotted in grey are the temperature and wind speed values corresponding to rain above the 75th quantile, and once again, the resulting hazard curve is shown in red with three different confidence intervals. The confidence intervals have been found using 1000 boot-strapped samples.

Some of the rainfall values in this data set seem unreasonably large, which may be due to some sort of measurement or recording error. For this reason, it would be unwise to draw any conclusions from these plots, although the data is sufficient to demonstrate the methods we are discussing.

As before, the methods can also be carried out for other event frequencies, such as 1 in 10,000 year events, and it is also straightforward to change the quantile of the conditioning variable that is used, depending on the situation of interest.
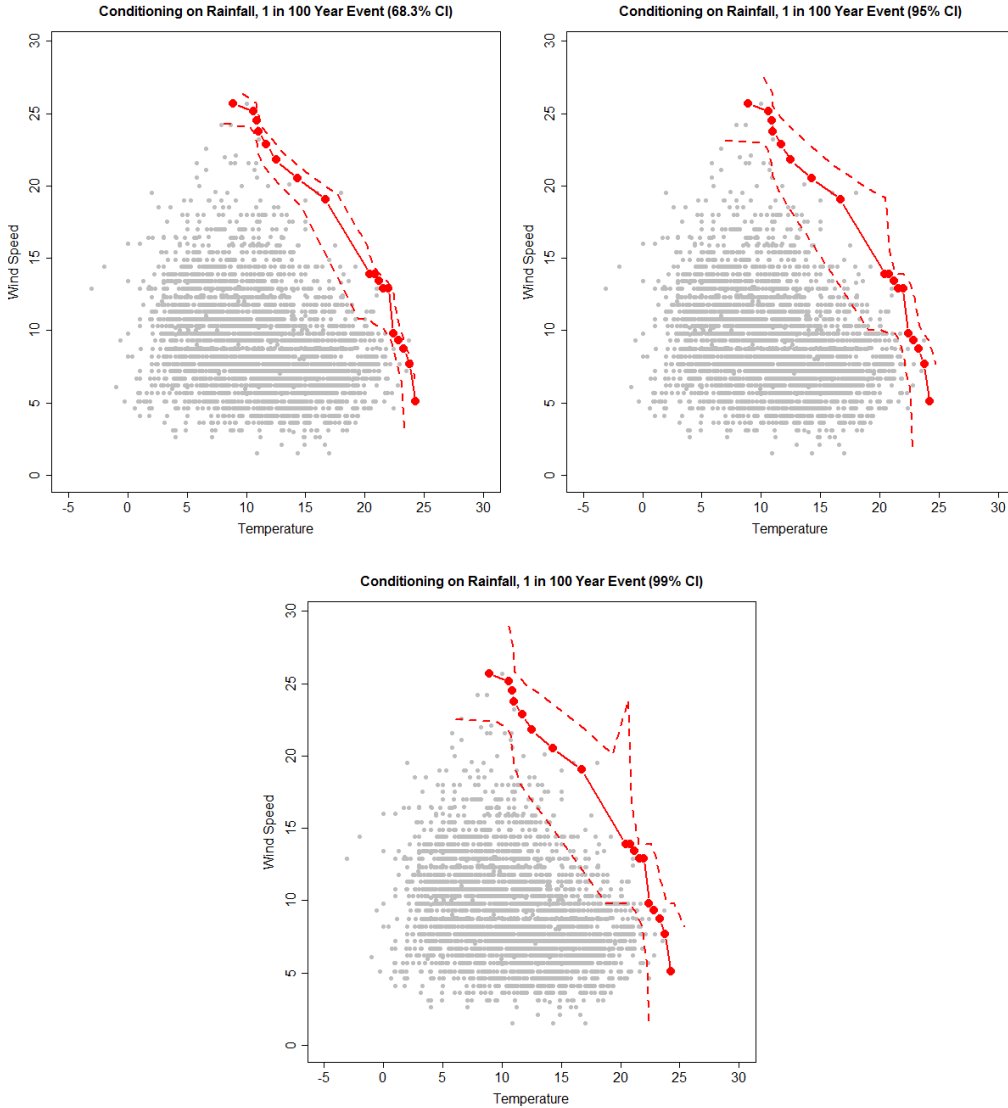
Figure 16: Hazard curve and confidence regions for maximum temperature and wind speed conditioning on high rainfall for 1 in 100 year events.

As was also mentioned in Section 3, an alternative method to use here would be to simulate data using Heffernan and Tawn's method conditioning on the fact that Z is greater than z, then using this to find the hazard curves. It is not clear which of these approaches will give the best results, so it is worth bearing in mind that there is an alternative when implementing either one of these methods. The most appropriate method may depend on the size of the data set.

5   Issues and Possible Future Work

There are several other areas that may be of interest to consider when constructing hazard curves such as those described in this document. In this section, we will outline some of these issues, and potential matters that may require further consideration if these methods are to be implemented in the future.

5.1 Independence Assumption

The first important issue to note is that the methods described here to calculate bivariate hazard curves assume that there is independence between different observations in each of the variables considered. In some applications this may not be an issue, but in the case of variables relating to weather, it may be too strong an assumption to make.

For instance, extreme events involving wind speed, temperature and rainfall often do not just occur on single days, and are instead frequently observed over multiple consecutive days. For instance, there may be a heatwave or storm event that lasts for several consecutive days, and leads to extreme observations of one or more of these variables on each of these days. In this situation, observations on different days cannot be said to be independent. As such, care should be taken when applying the methods described in this document, and ways to overcome the problems associated with this independence assumption is something that should be investigated further in any future work in this area.

5.2 Alternative Definition of a Hazard Curve

In this document, we have considered a bivariate hazard curve to be defined as a collection of values (x, y) satisfying

$$\Pr(X > x, Y > y) = p$$

for a specified value of p. However, this is not the only way that a hazard curve could be defined, as there is more than one way in which to define an extreme event.

One alternative is to consider values of (x, y) that satisfy

$$\Pr(X > x \text{ or } Y > y) = p,$$

again for some specified value of p, that may, for instance, correspond to a 1 in 100 year or 1 in 10,000 year event. This definition considers extreme events to be those where at least one of the variables is large.

To find values (x, y) satisfying this equation, we could use the fact that

$$\Pr(X > x \text{ or } Y > y) = \Pr(X > x) + \Pr(Y > y) - \Pr(X > x, Y > y),$$

so our aim is to find values (x, y) satisfying

$$\Pr(X > x) + \Pr(Y > y) - \Pr(X > x, Y > y) = p.$$

In the previous method, we chose to fix one of the values (x or y) at a quantile of the distribution for the corresponding variable, and the same approach may be taken here. Suppose we fix x to be x_q, denoting the qth quantile of the distribution of X. Then our equation becomes

$$(1 - q) + \Pr(Y > y) - \Pr(X > x_q, Y > y) = p,$$

or equivalently,

$$\Pr(X > x_q, Y > y) - \Pr(Y > y) = 1 - p - q.$$

It should once again be possible to use Heffernan and Tawn's method to estimate the value of y that satisfies this equation for given values of p and q.

For a fixed value of our probability, p, this equation could be solved for a variety of values of q. This would allow us to obtain several pairs of values of (x, y) satisfying $\Pr(X > x \text{ or } Y > y) = p$, which can be joined to form a hazard curve, as was the case previously. As before, we may instead choose to fix the value of y to be the qth quantile of the distribution of the variable Y, and find values of x satisfying

$$\Pr(X > x, Y > y_q) - \Pr(X > x) = 1 - p - q,$$

providing further points that can be used to construct the hazard curve. Uncertainty may once again be added to curves such as this using the bootstrapping method that was outlined in Section 4.2.

Although this approach is not demonstrated here, it is worth considering the fact that hazard curves may be defined in different ways, as one definition may be more useful than another for a given situation. It is important to consider the question that needs to be answered when deciding on the type of hazard curve that may be appropriate to use.
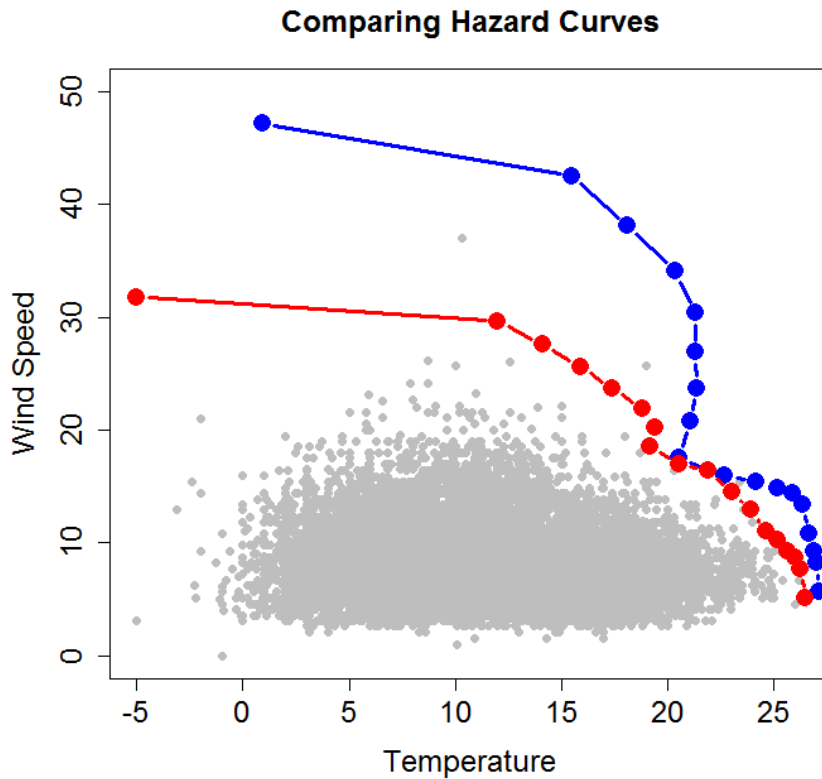
5.3 Ledford and Tawn

Figure 17: 1 in 100 year (red) and 1 in 10,000 year (blue) hazard curves for temperature and wind speed.

One noticeable characteristic of some of the hazard curves found in this document, particularly in Sections 4.1 and 4.2, is a sort of 'dip' in the middle of the curve that we may not expect to see. This is demonstrated in Figure 17, where a 1 in 100 year hazard curve is plotted with a 1 in 10,000 year hazard curve for temperature and wind speed data.

This feature may suggest that Heffernan and Tawn's method is not performing as well as hoped for this particular data set. This is most likely caused by the fact that the data is negatively dependent, i.e. there is a tendency for large values of one variable to occur with small values of the other variable. This negative dependence means that the step in Heffernan and Tawn's method where we simulate wind speed and temperature data given that wind speed is above y m/s or temperature is above $x°$C, is not producing values in the top right hand corner as often as it should.

This issue is demonstrated in Figure 18, which outlines why we are unable to simulate values in the top right hand corner. When we condition on temperature being above $x°$C, our simulated data appears to have a direction almost parallel to, or moving towards, the x-axis. When we simulate data conditioning on the fact that wind speed is above y m/s there is an improvement, but the set of simulated data is still not reaching very far into the top right hand corner, even for a large sample of 10,000 points. The fact that the 1 in 100 year and 1 in 10,000 year hazard curves overlap at around (20.5,17.5) in Figure 17 shows that this feature of the method is a concern, as we would not expect the hazard curves to intersect for two such different frequencies. This is an issue that should be investigated further.
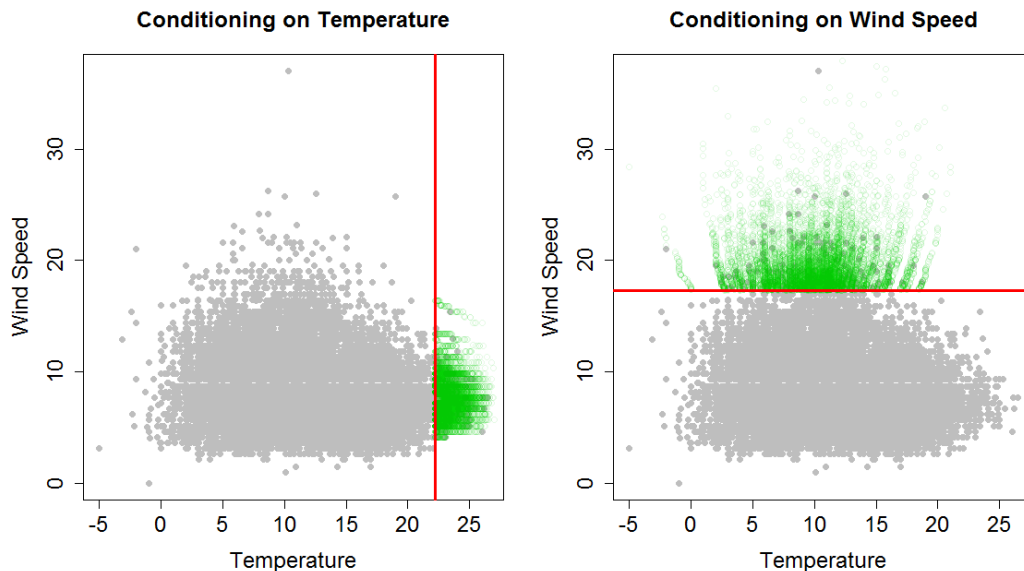

Figure 18: An illustration of simulating data using Heffernan and Tawn's method.

One consideration that may help here is to use the method of Ledford and Tawn (1996). As with Heffernan and Tawn's method, and other methods in multivariate extreme value theory, Ledford and Tawn's method requires transformation of the variables to a common distribution.

21

In this case, it is common to use the exponential distribution. For the variable X with distribution function F_X, transformation to the exponential distribution can be achieved via the equation

$$X_E = -\log(1 - F_X(X)),$$

and we can obtain Y_E, the transformation of variable Y to the standard exponential distribution, in a similar way.

In its simplest form, Ledford and Tawn's approach can be used to estimate probabilities of the form $\Pr(X_E > x, Y_E > x)$, i.e. the probability that both variables exceed the same quantile. This is useful in our situation since this is one of the cases in which our current approach, using Heffernan and Tawn's method, appears to fail.

Ledford and Tawn's model is of the form

$$\Pr(X_E > s, Y_E > s) = \mathcal{L}(e^s)e^{-\frac{s}{\eta}}, \tag{2}$$

where $\eta \in (0, 1]$, and $\mathcal{L}(.)$ is a slowly varying function at infinity, i.e. $\mathcal{L}(st)/\mathcal{L}(t) \to 1$ as $t \to \infty$ for all fixed $s > 0$.

In practice, we can estimate the value of the parameter η using the Hill estimator (Hill (1975)). Suppose we define the variable $M = \min(X_E, Y_E)$, and that we have realisations $m_1, ..., m_n$ of this variable. We denote by $m_{(1)}, ..., m_{(n)}$ the ordered realisations, where $m_{(1)} > m_{(2)} > ... > m_{(n)}$. We then fix a high threshold u, and suppose there are n_u realisations of M above this value. Then the Hill estimate of η, denoted by $\hat{\eta}$, is calculated using the equation

$$\hat{\eta} = \frac{1}{n_u} \sum_{i=1}^{n_u} \left(m_{(i)} - u\right).$$

More information on the Hill estimator is given in Section 9.5.2 of Beirlant et al. (2004), along with a discussion of alternative estimates of η.

By Equation (2), we have

$$\Pr(X_E > s + t, Y_E > s + t) = \mathcal{L}(e^{s+t})e^{-\frac{s+t}{\eta}},$$

from which it can be deduced that

$$\Pr(X_E > s + t, Y_E > s + t) \sim e^{-\frac{t}{\eta}}\Pr(X_E > s, Y_E > s).$$

The probability $\Pr(X_E > s, Y_E > s)$ can be estimated empirically for some value of x, and using the Hill estimate of η allows estimation of $\Pr(X_E > s + t, Y_E > s + t)$.

To put this in the context of constructing a bivariate hazard curve, we are interested in finding a value $s^\star$ such that

$$\Pr(X_E > s^\star, Y_E > s^\star) = p$$

for some specified value of p. Once we have chosen a value of s, this can be achieved using the equation

$$s^\star = s - \hat{\eta}\log\left(\frac{p}{\Pr(X_E > s, Y_E > s)}\right).$$

22

Once the value of $s^\star$ has been calculated, we can transform back to the original margins to find a point (x, y) satisfying

$$\Pr(X > x, Y > y) = p.$$

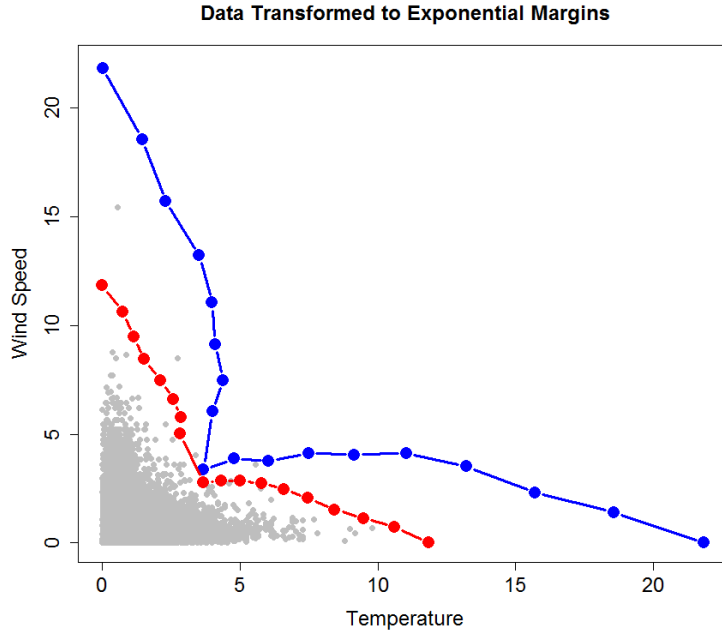**Data Transformed to Exponential Margins**

Figure 19: Example of transforming data to exponential margins with 1 in 100 year (red) and 1 in 10,000 year (blue) hazard curves.

Figure 19 shows the temperature and wind speed data transformed to exponential margins with the 1 in 100 year and 1 in 10,000 year hazard curves in these margins given in red and blue, respectively.

Using Ledford and Tawn's method, we estimate the point (3.8,3.8) in exponential margins to be exceeded once every 100 years, and we estimate the point (5.4,5.4) to be exceeded once every 10,000 years. As with Heffernan and Tawn's approach, we can use bootstrapping to get an idea of the sampling variability in the estimate of these points. Figure 20 shows the 1 in 100 year and 1 in 10,000 year hazard curves on exponential margins calculated using Heffernan and Tawn's method. The points in green on each plot represent 1000 bootstrapped estimates of the point $(s^\star, s^\star)$ satisfying $\Pr(X_E > s^\star, Y_E > s^\star) = p$ for $p = \frac{1}{365} \cdot 10^{-2}$ and $p = \frac{1}{365} \cdot 10^{-4}$, using Ledford and Tawn's method. The black lines on these bootstrapped points represent a 95% confidence interval for $s^\star$.

It is clear from these plots that Ledford and Tawn's and Heffernan and Tawn's approaches do not agree in these cases, since the confidence interval does not overlap the hazard curve in either of the plots. The situation appears to be worse in the 1 in 10,000 year hazard curve, where the fact that we are having to extrapolate further beyond the data than in the 1 in 100 case is having a negative impact on the estimation of the hazard curve. This is reiterated in the plot in Figure 21, which shows the same procedure carried out for 1 in 10 year events. In this

23

case, the confidence interval for the values from Ledford and Tawn's method overlaps the hazard curve calculated using Heffernan and Tawn's method. In this example, there is less extrapolation involved in the calculations.
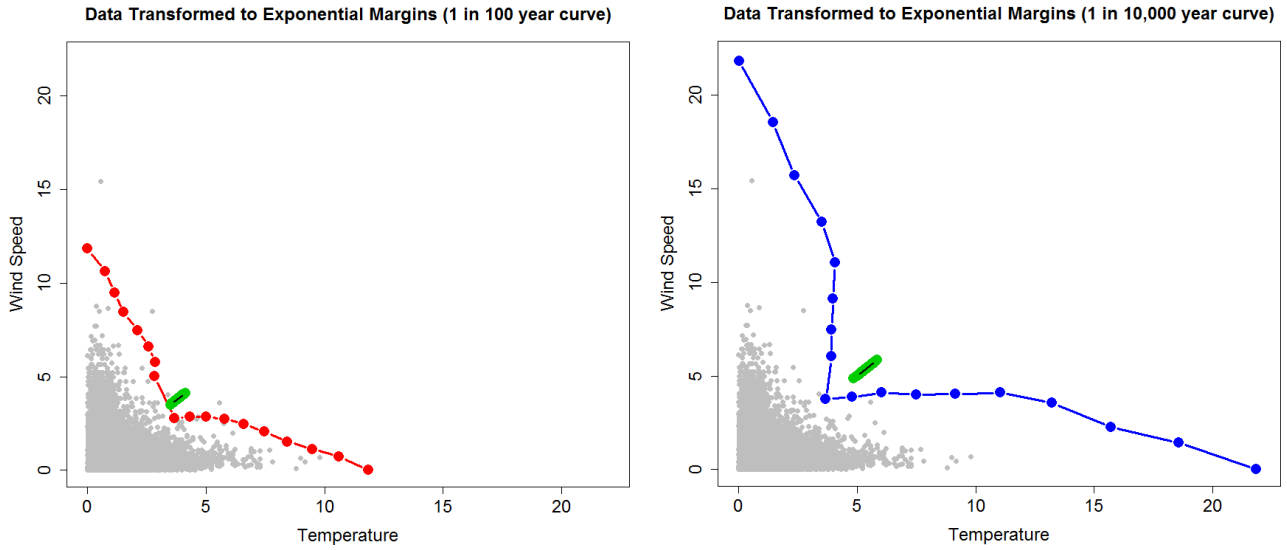

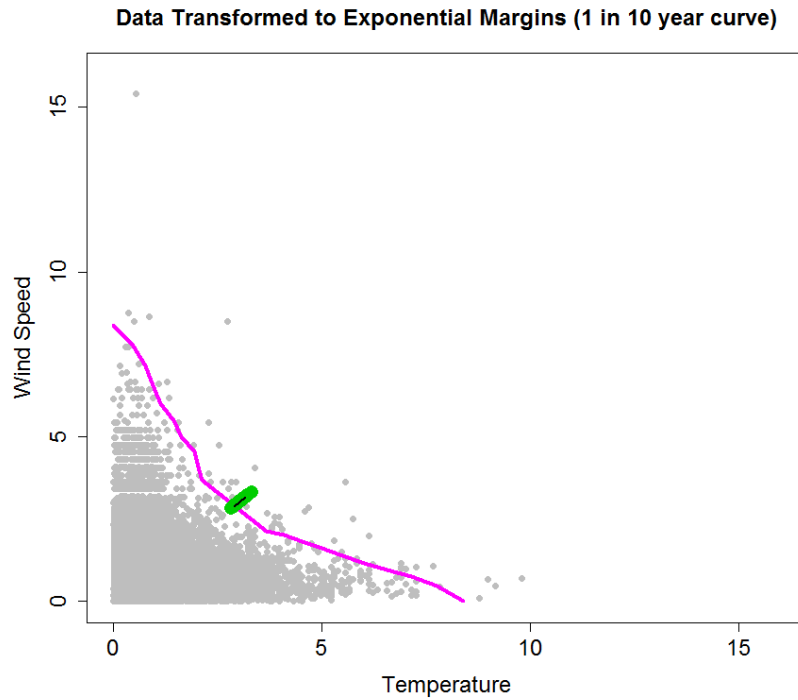Figure 20: Example of bootstrapped estimates from Ledford and Tawn's method.

Figure 21: Example of bootstrapped estimates from Ledford and Tawn's method for 1 in 10 year events.

The discussion presented here involving Ledford and Tawn's approach shows a possible problem with using Heffernan and Tawn's method to find a bivariate hazard curve when there is negative dependence in the data, particularly when this requires extrapolation for very rare events. It does, however, also present a possible solution to this problem. It may be worth investigating the possibility of using Heffernan and Tawn's method and Ledford and Tawn's method in conjunction with one another to create more reliable hazard curves.

6   Conclusion

In this report, we have introduced some techniques in univariate and multivariate extreme value theory and presented a method to create bivariate hazard curves and assess the uncertainty associated with them. As well as considering events where values are exceeded in two variables, we discussed creating hazard curves for combinations of low values and high values. We also considered cases where more than two variables are of interest, in particular focussing on creating these hazard curves when conditioning on one of the variables being large.

We also discussed some of the issues with this method, and considerations that should be taken into account when applying these methods. In particular, we discussed the fact independence assumption that must be made when applying the methods used to create the hazard curves, and the fact that it is possible to find alternative ways of defining a hazard curve due to the different ways in which we can define an extreme event. Finally, we considered the issue of negative dependence and extrapolation, introducing the possibility of using Ledford and Tawn's method alongside that of Heffernan and Tawn to create better estimates for the hazard curves.

References

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D., and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Lecture Notes in Control and Information Sciences. Springer.

Coles, S. G., Heffernan, J., and Tawn, J. A. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Southworth, H. and Heffernan, J. E. (2013). *texmex: Statistical modelling of extreme values*. R package version 2.1.