Honours Research Project on

Modelling extreme events using point processes

Alice Thesen & Manon Shapiro

THSALI002 SHPMAN002

Supervisor: Dr Birgit Erni

Department of Statistical Sciences

University of Cape Town

October 24, 2022

Abstract

Temporal data consisting of environmental extremes presents a particular challenge in the modelling process, due to its capacity to occur in clusters, hence exhibiting dependence between events. This analysis aims to assess the suitability of using point process models and extensions to model environmental time series of extreme events. Point process models have frequently been applied to seismological and, more recently, financial data, but have seldom been applied to a more general environmental time series of extremes. The specific data set used was provided by SAEON and consists of CO_2 mmol/mol readings taken at fixed intervals from a Savanna site in Kimberly, South Africa. This project explores mechanisms of determining thresholds aimed at capturing the extremes, which were then used to extract the data in the required point process form. The models formulated and compared in this project were the Poisson process and two extensions thereof. The first extension is known as the Hawkes process and aims to account for the clustering of events by including the history of events when calculating the conditional intensity of the process. The second extension is a variation of the Hawkes process which includes additional information about the event of interest, known as marks, in the conditional intensity formulation.

The approach utilised to account for the seasonality observed in the data set was to split the data into the three complete observed seasons, determining a suitable threshold for each and modelling the extremes of each season separately. It was found that the Poisson process model was a very poor fit for the data, as it failed to model the clustering of the extremes. The Hawkes process model captured the clustering in the data relatively well, particularly for the less tightly clustered events of the wet season. For the dry seasons, the Hawkes demonstrated improved fit when compared to the Poisson process, but still exhibited a tendency to overestimate the inter-arrival times between events, indicating that this model formulation struggles to capture very tight temporally clustered events. The marked Hawkes process model showed very little to no improvement on the Hawkes models produced for each season, indicating that the inclusion of a marked parameter is an unnecessary added complication when aiming to model this system. Overall, it was found that point process models may be very useful when modelling time series of environmental extremes. In particular, the Hawkes process was found to capture clustered data well, provided the observed temporal clustering is not too extreme.

Acknowledgements

Thank you to our supervisor, Dr Birgit Erni, for all of the well-thought-out advice, the constructive and thought-provoking discussions and for always having a calming presence.

Additionally, thank you to Dr Etienne Pienaar and Assoc. Prof. Tim Gebbie for lending their time and expertise to assist us with our project.

A special thanks to our parents for their unconditional support, love, encouragement and snacks.

Contents

List of Figures

List of Tables

1 Introduction

Understanding the mechanisms of occurrences of extreme events is useful in a multitude of different fields, from financial markets to epidemiology. Modelling time series data of environmental extremes can provide some insight into the underlying process which results in the occurrences of such events and how these events change over time. There are well-defined and established schools of thought for evaluating distributions of extreme readings, such as Extreme Value Theory, but these often make the assumption of independence between events. This is usually not the case due to the clustering nature of extreme events. For example, consider the aftershocks that follow earthquakes, which can trigger further earthquakes. In this way, one earthquake can trigger another. It is clear that events in this system cannot be considered unrelated to one another when occurring in a given time period.

This project is concerned with the application and suitability of point processes to model environmental time series of extremes. A point process model produces a conditional intensity function, which defines the expected number of events of interest at a given time in an interval. Initially, the basic Poisson process will be explored, which considers the event of interest as a Poisson-distributed random variable. A further objective is to investigate the suitability of applying some extensions of point processes to time series data of environmental extremes, such as the self-exciting Hawkes process and marked Hawkes process. Hawkes processes capture temporally clustered data well, as they take into account the history of events when generating the conditional intensity function. Marked data gives further, relevant information about the events of interest which may enable more accurate modelling of the system.

The data used in this analysis consists of climatic CO_2 mmol/mol readings recorded from a savanna site in Kimberly, South Africa. The fluctuation and exchange of climatic CO_2 is a fundamental component of any natural ecosystem. Savannas are a widespread biome that can be found globally. On the African continent, tropical grasslands and savannas cover more land than tropical forests (Surlock and Hall, 1998). Modelling the climatic CO_2 extremes that occur in grasslands may provide some insight into the mechanism of the systems that produce extreme readings in these biomes.

This project will be structured as follows: Section 2 explores existing literature identifying various applications of point process models and extensions. The model formulations, likelihoods, compensators and other relevant formulae and theorems will be given and discussed in more depth in section 3. Section 4 gives a description of the data, the data collection site and data management procedures. Section 5 provides an exploration of the data set, investigating the extremes for each season in the analysis. Sections 6, 7 and 8 consist of model formulation and analyses for the first dry season extremes, the wet season extremes and the second dry season extremes, respectively. Each section provides the respective Poisson, Hawkes and marked Hawkes formulations and intensity plots, as well as model-fitting assessment procedures for the models of each season's extremes. This is followed by section 9, giving a detailed discussion and interpretation of the results found in sections 6, 7 and 8. Section 10 concludes with key findings and future considerations.

2 Literature review

Modelling extreme events often requires modelling the tail distribution of data, while most common methods of modelling instead focus on modelling the central bulk of the distribution. An added complication arises when these extreme events occur over time as there is often clustering and dependence among events. Methods that have been used to approach the issue of modelling extreme events are using Extreme Value Theory (EVT), a point process or a combination of the aforementioned methods.

EVT is a mechanism of modelling and describing the distribution of extreme events. It is well established in many different fields such as engineering, insurance and science (see e.g. Reiss and Thomas, 1997). It has been used to analyse financial markets, particularly in the area of risk management, as it is the extreme probabilities and quantiles that are of interest (see e.g. Diebold, Schuermann, and Stroughair, 1998; Gilli and Kellezi, 2006). EVT is an attractive option when the object of interest is the tail of the distribution, rather than the centre, as the relevant estimation methods are designed for this (Diebold, Schuermann, and Stroughair, 1998).

However, EVT has the pitfall of assuming that the data is independent and identically distributed (i.i.d.), despite these extreme values often occurring in clusters. To overcome dependence between events, generalisations that remove the clustering condition have been utilised, such as the generalised extreme-value (GEV) distribution. Nonetheless, these generalisations are only able to model the extreme value distribution over time and do not deal with how or when these events occur. Hence a model is needed that takes into account the history of all the previous events and the dependence between them.

The alternative method for modelling dependent and clustered extreme values is to use a Point process, which was originally explored by Cox and Lewis (1996), Cox and Isham (1980) and Barlett (1963a; 1963b). Point process have many applications and consists of a simple Poisson process, Hawkes processes and various other extensions. Point processes have been used to model extreme values and have been applied in various fields. Boano-Danquah, Sigauke and Kyei (2020) analysed extreme peak loads in South Africa using a point process, Fuest (2009) modelled financial time series extremes and Descombes and Zerubia (2002) used point processes for image analysis and processing.

Similarly, the Hawkes processes and extensions have been applied to many fields, mostly finance (see e.g. Hardiman, Bercot and Bouchaud, 2013; Lapham, 2014) and social networks (see e.g. Lukasik et al., 2016; Nie et al., 2020), although it has been used in other applications such as neuroscience, medicine and crime prediction (Hawkes, 2018). Recently, the Hawkes process has been used to model the spread of COVID-19 (see e.g. Garetto, Leonardi and Torrisi, 2021; Chiang, Liu and Mohler, 2022).

In terms of environmental time series, the point process models have many applications. The Poisson process model has been used to analyse and model extreme seasonal rainfall events (see e.g. Ngailo et al., 2016; Morales and Rodrigues, 2022) as well as to model extreme

weather events over Germany (Dalelane and Deutschlander, 2013). EVT has been used in conjunction with a Poisson process to model the ground-level ozone over 13 years in Houston, Texas (Smith, 1989). Similarly, Koh et al. (2021) used a spatiotemporal point process to model wildfire intensity and size by using both EVT and point processes. An example of a point process extension is the marked point process model which has been used to model raindrop-size distributions over time (Smith, 1993) and used for forest studies where both time and spatial elements are present (Gavrikov and Stoyan, 1995).

Similarly, the Hawkes processes have been used for different environmental extremes, such as seismology, where both time and location play an important role (Ogata, 1988; Fox, Schoenberg and Gordon, 2016). A similar model can be used to model invasive plant and animal species like the red banana plants which are invasive in Costa Rica (Balderama, Schoenberg, Murray and Rundel, 2012), thereafter an optimal intervention plan can be obtained as seen in Gupta et al. (2018). Although there have been some applications of the Hawkes process to ecological and environmental times series, the most common application of the Hawkes process is to model financial times series.

Often environmental time series consist of readings taken at fixed intervals and are not in the required point process form. Often, the extreme occurrences of some event are of interest. Therefore a threshold above which events are considered extreme is required. EVT can aid in finding the threshold value as there are numerous tools and diagnostic plots available within the EVT framework to aid in the choice of a threshold. Boano-Danquah, Sigauke and Kyei (2020) use extremal mixture models from EVT, however, the process needs to be homogeneous and as such the data is stationarised. Threshold stability plots can be used to identify a suitable threshold (see e.g. Bommier, 2014, pg 21; Koh, 2021, pg 7) or the mean residual life plot (see e.g. Scarrott and MacDonald, 2012; Lapham, 2014).

Dalelane and Deutschländer (2013) employed a different approach and used three different thresholds in their modelling; the 0.9-quantile, the 0.95-quantile and the 0.99-quantile in an effort to explore a range of thresholds when making this important decision. They modelled the Poisson point process, using the aforementioned threshold, as it is expected that when the threshold is high enough, the events will become independent and follow the Poisson distribution. Similarly, Ngailo et al. (2016) use the 0.99-quantile of the distribution above which the events are considered extreme.

Hawkes (1971) introduced the simple Hawkes point process model which specifies a self-exciting relationship between clustered events. Hawkes produced a class of theoretical models that calculate the intensity of the current event as a result of the history of previous events. These are stochastic point processes known as self-exciting and mutually exciting. Hawkes originally suggests using an exponentially decaying function for the kernel, which is a special case as it simplifies the derivations and ensures a feasible analytic solution. Using an exponentially decaying kernel is the most common approach (see e.g. Oakes, 1975; Liniger, 2009; Dassios and Zhao, 2013)

An important variation of the Hawkes process can be understood by using a Poisson cluster process, which was proved to be an equivalent representation by Hawkes and Oakes (1974). This interpretation makes use of an immigrant-birth representation whereby point events can be separated into two types; an immigrant or offspring point (see Daley and Vere-Jones, 2003).

Another alternative technique to model the Hawkes process is using a Recurrent Neural Network (see e.g. Mei and Eisner, 2017; Zhang, Lipani, Kirnap and Yilmaz, 2020; Malaviya, 2021). These Recurrent Neural Network models all performed well, but require extensive knowledge regarding Neural Networks and therefore the modelling process can be challenging.

A formulation of the Hawkes process that is not covered in literature often is the self-regulating case of the Hawkes process which accounts for both excitation and potential inhibition effects. Bonnet, Herrera and Sangnier (2021) explored an approach which can handle both effects. This model is advantageous when the presence of an event decreases the chance of another happening, a characteristic often observed in neuroscience. Importantly, the cluster representation is no longer valid when both excitation and inhibition effects are present.

There are many Hawkes extensions that can be used to further capture information available in the system. Although, there is limited literature which utilises and compares both the Hawkes process and Hawkes extensions in the same paper. It is more common to have a paper either focusing on the Hawkes process or an extension.

A common extension is that marks can be included to more accurately capture the conditional intensity, where a mark contains additional information about the events, such as the location or the magnitude. Marks are often used in financial applications where the conditional mark distribution is a generalisation of the generalised Pareto distribution (GDP) (McNeil, Frey and Embrechts, 2005). The marks are included in the intensity formulation through an exponential impact function. Lapham (2014) explores various conditional mark distributions, initially starting with the GDP and subsequently letting various parameters equal zero to create different models. There was no clear variation of the marked Hawkes process that performed best across all the criteria considered.

Originally Ogata (1988) created the temporal Epidemic Type Aftershock-Sequence (ETAS), an extension of the Hakwes process model, which only included the effect of magnitude of an earthquake and failed to use a spatial factor. This popularised the power law, which can be used for the decay component of the kernel function. Thereafter, Ogata (1998) extended the popular ETAS model to become more flexible to capture the spatiotemporal interactions between earthquake occurrences. This marked point process used both magnitude and coordinates as the chosen marks, rather than just the magnitude. The ETAS model has subsequently been applied in ecology, epidemiology, crime prediction and gang warfare (Hawkes, 2018). A similar spatiotemporal point process model was used to model wildfire intensity where the mark is the location relating to the burnt areas (Koh et al, 2021).

As mentioned earlier, another extension of the self-exciting Hawkes process is to model multivariate mutually exciting extreme events, whereby the multivariate extreme exceedances are treated as realisations of the univariate point process (Hawkes, 1971; Li and Zha, 2015). This property is advantageous for maximum likelihood estimation and has been used by Embrechts, Liniger and Lin (2011) and Grothe as well as Korniichuk and Manner (2014) to model events observed in a financial time series. The clusters present are due to both the occurrence time and magnitude mark of the extreme exceedance above some threshold as well as another influential variable, hence Grothe, Korniichuk and Manner (2014) use a multivariate marked Hawkes process.

For estimation of the Hawkes and marked Hawkes parameters, there are various techniques that can be used. Koh et al. (2021) and Holden, Sannan and Bungum (2003) used a Bayesian method for estimating the parameters. However, this requires a Bayesian prior for each parameter to be specified as well as a posterior based on observed data.

Another parameter estimating option is to use methods of moments, which Grothe, Korniichuk and Manner (2014) found inferior to using maximum likelihood estimation when modelling multivariate mutually exciting extreme events. Using the direct numerical maximisation (DNM) of the likelihood is the most common route to finding the parameters. Ogata (1988), Grothe, Korniichuk and Manner (2014), and Lapham (2014), among others, found the parameter estimates by maximising the log-likelihood.

Lapham (2014) also explored using an expectation-maximisation (EM) algorithm, which utilises the Poisson cluster process interpretation and found that for the simple Hawkes process, using DNM is preferred. Moreover, the EM algorithm can be unstable and computationally intensive (Veen and Schoenberg, 2008).

Issues can arise when using MLE to determine the model parameters. These issues are with regard to computational efficiency and multiple maxima in the likelihood (Daley and Vere-Jones, 2003; Guo and Luk, 2013; Lapham, 2014). Computational difficulties arise as the evaluation of the log-likelihood requires repeated evaluation of the conditional intensity function. This computing process is made more efficient by using a recursive formula as presented by Ogata (1981). The recursive formula can be constructed due to the Markov property of the intensity, which is maintained when an exponential decay and impact function is used. Moreover, both the Hawkes and marked Hawkes processes have a complicated log-likelihood function which is often non-convex. This non-convex log-likelihood can have many local maxima, resulting in the log-likelihood not converging towards the global maximum. Lapham (2014) suggests using a range of starting values to identify the global maximum.

Once the parameters have been estimated and the conditional intensity is found, the Hawkes processes and Hawkes extensions models need to be checked and compared. To determine which variation of the model should be selected, the maximum likelihood associated with the estimated parameters is used and the models are subsequently compared using Akaike's Information Criterion (AIC; Akaike, 1974). In literature, this is the most common measure

for model selection. Examples of AIC being used include the work of Wang, Bebbington and Harte (2012), Lapham (2014), Ogata (1988) and Ogata and Zhuang (2006).

In terms of goodness-of-fit tests, once the best model has been selected, the most common approach is through visual diagnostics (Koh et al, 2021; Lapham, 2014; Laub, Lee and Taimre, 2021). First, event inter-arrival times need to be simulated from the fitted model. One such simulation technique is a simulation by thinning in order to get multiple inter-arrival times from the proposed model. These inter-arrival times are reviewed using a quantile-quantile plot (Q-Q plot). The counting process of the event arrivals is compared to the first moment of that process. Lastly, the random time change theorem for goodness-of-fit is used for the temporal component and graphically, the goodness-of-fit is constructed from the 'residual process' as described by Ogata (1988). Examples of goodness-of-fit tests can be found in the works of Ogata (1988), Liniger (2009), Lapham (2014), and Bonnet, Herrera and Sangnier (2021).

With regards to environmental time series, there tends to be seasonality present due to climate factors varying with changing weather conditions. These weather conditions can create environmental clustering of events within a system. There are various factors that affect the CO_2 level in the system. Firstly, wind direction and speed are identified by Al-Bayati et al. (2020) as having a noticeable effect on the concentration of CO_2 in Iraq.

Patterns of rainfall have been identified to have a large influence on the effect of CO_2 in southeastern Tasmania, Australia (Hovenden, Newton and Wills, 2014). Summer rainfall has a stimulating effect whereas autumn and spring have a inhibiting effect on the CO_2 response, indicating there is a seasonal variation of effect of rainfall. This is supported by Räsänen et al. (2017) who found that at a grazed savanna grassland in Welgegund, South Africa, there was high inter-annual variation in the CO_2 flux during the wet season while the magnitude of CO_2 fluxes was similar in the dry season. Although there is an effect, light rainfall of about 0.6mm is found to only have a profound short-term effect on CO_2 (Kim, Verma, and Clement, 1992; Zepp, et al., 1996) As there is seasonal variation, the CO_2 values considered extreme may vary throughout the year.

Similarly, Zepp et al. (1996) found that soil temperature and moisture are found to be crucial climate factors that influence soil respiration of CO_2 into the atmosphere. Soil moisture content exerts a strong control on CO_2 fluxes (Williams, et al., 2009). Moreover, soil moisture and soil temperature partly control carbon fluxes during the wet season whereas, in the dry season, carbon is less sensitive to soil temperature and moisture (Räsänen et al. 2017). Indicating that environmental factors can have a different effect on CO_2 during the different seasons.

There is also a link present between precipitation and soil moisture as soil CO_2 flux tends to be at a peak when there are high values of extractable soil water percentage, occurring when the soil water was replenished by precipitation (Kim, Verma, and Clement, 1992). Lastly, soil water availability, as well as air temperature, are primary factors that control CO_2 perspiring

from the soil in temperature grasslands (Kim, Verma, and Clement, 1992). Air temperature appears to moderate daytime CO_2 flux, but this occurs when a site is wetter. As for some sites from Scanlon and Albertson (2004) study, there was no observed temperature effect on CO_2 flux.

Therefore the varying seasonal level of CO_2 for the wet and dry seasons should be considered when modelling. There are various ways to account for this. Boano-Danquah, Sigauke and Kyei (2020) grouped the electrical demand into the four seasons present during the year as the demand is known to vary significantly throughout the year. Thereafter they found a season-specific threshold which can be seen as a time-varying threshold. This method allowed for each season's extremes to be accurately modelled.

As an alternative to varying the threshold, covariates can be included to capture the non-linear influence they play in the level of CO_2. Koh et al. (2021) used covariates through component-specific smooth functions which vary with the season and were within a Bayesian spatiotemporal model. Whereas Morales and Rodrigues (2022) opted to include a seasonal cycle component in the intensity function when using a Poisson process to model rainfall. This enables control of possible effects produced by event occurrences in regular periods.

3 Statistical Methods

This section describes the theory of counting processes and point processes. As the aim of this analysis is to model extremes, a threshold must be set beyond which events are considered extreme. Initially, this section will deal with the methodology behind threshold selection before defining counting and point processes. The extensions of the point processes that will be generalised to are the Poisson, Hawkes and marked Hawkes processes. The material, theorems, proofs and derivations in this section are largely based on those of Laub, Lee and Taimre (2021), who provide a thorough treatment of the various point processes and their general theory. The following will be modelled using `R` Core Team (2022). All of the log-likelihoods, conditional intensity functions, compensators, moments, simulations and methods for model diagnostics were coded using various point processes and their general theory. The following will be modelled without the use of any `R` packages.

3.1 Threshold determination

This project is concerned with the modelling of extremes. Thus, it is necessary to decide to find a reasonable threshold beyond which readings are considered extreme. The decision of a threshold is not straightforward and must be chosen carefully as there is a trade-off between bias and variance in parameter estimates (Boano-Danquah, Sigauke and Kyei, 2020). The threshold can be defined such that the population tail can be well approximated by an extreme value model. There are a few methods for selecting a threshold; Coles (2001) suggests using the mean residual life plot, while another method is using a rule of thumb.

Mean residual life plot

The mean residual life (MRL) plot, also known as the mean excess plot, uses the expectation of the Generalised Pareto Distribution (GPD) excesses to identify a suitable threshold.

The MRL plot visualises the mean exceedances beyond a variable threshold u by plotting the following points consisting of the threshold and the mean excess:

$$\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_i - u)\right),$$

where $x_1, \ldots, x_{n_u}$ are the n_u observations that exceed some threshold u. The mean excess for a given threshold is the average magnitude of the exceedances from that particular threshold. Confidence intervals can be found for the mean excess because the empirical mean is theoretically normally distributed. However, this normality does not hold for points where there are limited exceedances.

The threshold level is selected to be the lowest level whereby all the higher threshold mean excesses are consistent with a straight line such that all the exceedances follow a GDP (Gebbie, 2022). Although this method is from EVT, it can be used as a technique for finding a threshold in this analysis because, according to EVT, the rate of exceedances above a suitably high threshold should be a Poisson process such that the events follow a Poisson distribution (Smith, 1989; Scarrott and MacDonald, 2012).

In order to generate the MRL plot, two packages were used; `evmix` (Hu and Scarrott, 2018) and `POT` (Ribatet and Dutang, 2022). Coles (2001) does acknowledge that interpreting the MRL plot can be challenging. A Rule of Thumb for threshold determination is also used.

Rule of thumb

Frequently a simple 90th percentile is used but from a theoretical point of view, it is inappropriate so another method should be used (DuMouchel, 1983). There are two rules to be used:

$$\eta = \sqrt{n} \quad \text{and} \quad \eta = \frac{n^{\frac{2}{3}}}{\log(\log(n))}$$

where n is the number of observations and η is some threshold.

Ferreira, de Haan and Peng (2003), among others, use the first rule known as the square root rule while others use the second rule which was proposed by Loretan and Philips (1994).

A threshold can be used to extract data readings taken at fixed intervals into a form suitable for analysis by point processes and various extensions thereof. Before a point process can be defined, it is necessary to define a counting process.

Counting Process

A counting process is a stochastic process $(N(t) : t \geq 0)$ that can be thought of as a cumulative count of the number of events in a system up to a current time t. It can take on values that are integer and non-negative. It is a non-decreasing function which satisfies the following conditions:

- $N(0) = 0$
- $N(t) \geq 0$
- if $s < t$, then $N(s) \leq N(t)$
- Is a right-continuous step function with increments of $+1$.

The history of events up to time v can be denoted $\mathcal{H}(v)$

Point Process

A point process is a collection of points occurring in a mathematical space. A temporal point process occurs along a time axis for a set of observed time indices $\mathbf{T} = \{T_1, T_2, ...\}$, taking values in the range $[0, \infty)$. The time indices are the times at which the corresponding counting process $N(.)$ has jumped. If the number of points is finite and $\mathbb{P}(0 \leq T_1 \leq T_2 \leq ...) = 1$, then $\mathbf{T}$ is a simple point process.

A point process that only contains extreme events can be extracted from a dataset by using a threshold above which the events are considered extreme. Therefore if observations $x_1, ..., x_n$ represent data taken at some constant interval and suppose u denotes an arbitrary and sufficiently high threshold, then if $x_i : x_i > u$, those observations $x_1, ..., x_i$ are extreme events exceed some threshold u.

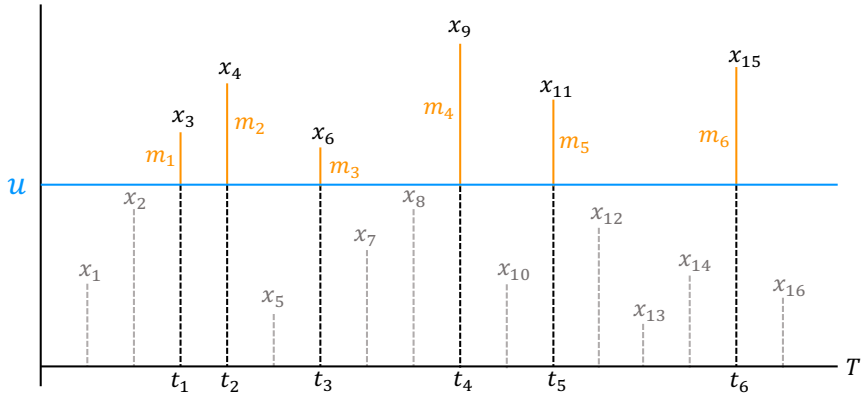

Figure 1: An illustration of a series of observations $x_1, x_2, ..., x_{16}$ with the chosen threshold u (blue). This threshold can be used to extract the marked point process realisation containing extreme events with their respective marks $(t_1, m_1), (t_2, m_2), \ldots, (t_6, m_6)$ from the observations.

An extension of a point process is the marked point process whereby marks can be included in the model formulation. An example of a mark is the magnitude of the events above the chosen threshold u such that the magnitude of exceedances can be found by $x_i - u$ for all $x_i > u$. Figure 1 illustrates the resulting set of observations and their corresponding marks $(t_1, m_1), (t_2, m_2), \ldots, (t_6, m_6)$ after only selecting the events above the threshold.

This extracted marked point process shows a discrete-time point process due to the events t being non-negative integers. However, this discrete-time point process will be treated and modelled as if it was a continuous-time process.

3.2 Poisson process

Homogeneous Poisson process

A counting process is a homogeneous Poisson process with rate λ if the following properties are met:

- The total number of points on the interval (I) is a Poisson distributed random variable with a rate parameter $\lambda|I|$ where $|I|$ is the duration of the interval.
- For any n disjoint intervals, $I_i, I_j, \ldots, I_n$, the random variables giving the number of events in each interval, denoted $N(I_i), N(I_j), \ldots, N(I_n)$, are independent.

These properties have the resulting consequence that the inter-arrival times between events are independent and exponentially distributed. The conditional intensity of a Poisson process is the rate at which events are expected to occur in the infinitesimal time interval around a current time. The conditional intensity modelled by a Poisson process is independent of the past and is constant at a value of λ. This can be expressed mathematically as

$$\mathbb{P}(\lambda|\mathcal{H}(v)) = \mathbb{P}(\lambda).$$

These characteristics can be used to assess the suitability of using a Poisson process to model a given data set. Quantile-quantile plots and graphs of the empirical versus theoretical distribution of the inter-arrival times are used to assess whether the inter-arrival times are exponentially distributed. Model validation is performed through the use of pseudo-residuals.

Inhomogeneous Poisson process

A counting process is an inhomogeneous Poisson process with a time-varying rate λ if:

- For any interval $I = (a, b]$, the number of events are Poisson distributed with rate parameter $\int_a^b \lambda(s)ds$
- For any k disjoint intervals, $I_i, I_j, \ldots, I_k$, the random variables $N(I_i), N(I_j), \ldots, N(I_n)$ are independent.

The consequence of these characteristics is that the inter-arrival times are no longer independent.

The cumulative distribution function of the next arrival, conditional on the history up until the last arrival v, $\mathcal{H}(v)$, can be shown to be:

$$F(t|\mathcal{H}(v)) = \int_v^t \mathbb{P}(T_{k+1} \, \epsilon \, [s, s+ds] \mid \mathcal{H}(v)) \;\; ds = \int_v^t f\left(s \mid \mathcal{H}(v)\right) ds, \tag{1}$$

where $f\left(s \mid \mathcal{H}(v)\right) ds$ is the probability density function of the next event given the history of events, $\mathcal{H}(v)$. This is called the conditional arrival distribution. The joint PDF of the event arrivals $t_1, t_2, ..., t_k$, can then be found by the chain rule to be:

$$f(t_1, t_2, ..., t_k) = f(t \mid \mathcal{H}(v)) = \prod_{i=1}^{k} f(\, t_i \mid \mathcal{H}(t_{i-1})). \tag{2}$$

For ease of notation, the cumulative distribution and joint probability density function shall be referred to as $F^*(t)$ and $f^*(t)$ respectively. It can be difficult to work with $f^*(t)$, so it can be beneficial to instead use the **conditional intensity function** which can be defined as:

$$\lambda(t) = \frac{f^*(t)}{1 - F^*(t)}, \tag{3}$$

which is equivalent to:

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{E}[N(t+h) - N(t) \mid \mathcal{H}(t)]}{h}. \tag{4}$$

The intensity function of a Poisson process can be thought of as the expected number of random events per unit length of the process, in the vicinity of the current location t.

The **compensator function** is the integrated conditional intensity function, which is often used for parameter estimation and model checking. It is given as:

$$\Lambda(t) = \int_0^t \lambda(s) \; ds. \tag{5}$$

By integrating the conditional intensity over some interval $[0, t]$, the compensator function gives the expected number of events in the interval $[0, t]$, or $\mathbb{E}[N(t)]$.

If we observe all of the arrival times over the interval $[0, T]$, denoted $t_1, \ldots, t_{\mathbf{T}}$, then the **likelihood for the point process** can be shown to be

$$L = \left[\prod_{i=1}^{N(T)} \lambda(t_i) \right] e^{-\Lambda(T)}, \tag{6}$$

where $N(T)$ is the total number of observed events over the interval $[0, T]$. Taking the natural logarithm, the log-likelihood is

$$\ell = \sum_{i=1}^{N(T)} \ln(\lambda(t_i)) - \Lambda(T). \tag{7}$$

The proof for equation (7) can be found in Laub, Lee and Taimre (2021, pg 38-40).

3.3 Hawkes process

A Hawkes process is an example of a non-homologous point process, meaning that the inter-arrival times of events are no longer considered independent. Hawkes processes take into account the entire history of events when modelling the conditional intensity function. The Hawkes conditional intensity is of the form

$$\lambda(t) = \lambda + \int_0^t \mu\ (t - v)\ dN(v), \tag{8}$$

where λ is the baseline intensity and μ is the excitation function. Using $t_1, t_2, ..., t_k$ to denote the sequence of past arrival times, the Hawkes intensity can also be given by:

$$\lambda(t) = \lambda + \sum_{i:t_i<t} \mu\ (t - t_i), \tag{9}$$

where $i \in \{0, k\}$.

Both λ and μ can be specified to suit the specific context of the data. A particular formulation of the excitation function that will be investigated is one that self-excites and decays exponentially. In this case, $\mu(t) = \alpha e^{-\beta t}$. The effect of an arrival on the system of self-exciting Poisson processes are that the conditional intensity is increased by an arrival and decays thereafter. The conditional intensity of a **Hawkes process with exponential decay** can be shown to be:

$$\lambda(t) = \lambda + \int_0^t \alpha e^{-\beta(t-s)}\ dN(s) = \lambda + \sum_{i:t_i<t} \alpha e^{-\beta(t-t_i)}. \tag{10}$$

When an arrival enters the system, it instantaneously increases the conditional intensity function by rate α. Over time, this arrival's influence on the conditional intensity decays

by rate β. It is important to note that self-exciting Poisson processes can cause temporal clustering of events over the observed interval, $[0, T]$. The parameters α and β must be chosen to avoid explosion, hence we incorporate an additional constraint that $\alpha < \beta$.

The Hawkes process with exponential decay also lends itself to use by its **Markov property**. Because the excitation function is exponentially decaying, the result is that the conditional intensity $\lambda(t)$ is also an exponentially decaying function (except at jump times). Imagine that we observe the first k arrivals and that we are looking at some future time s where $t_k < s < T_{k+1}$:

$$\begin{aligned}
\lambda(t) &= \lambda + \sum_{t_i<t} \alpha e^{-\beta(s-t_i)} \\
&= \lambda + \sum_{t_i\leq t_k} \alpha e^{-\beta(s-t_k+t_k-t_i)} \\
&= \lambda + e^{-\beta(s-t_k)} \sum_{t_i\leq t_k} \alpha e^{-\beta(t_k-t_i)} \\
&= \lambda + (\lambda(t_k) + \alpha - \lambda)e^{-\beta(s-t_k)}.
\end{aligned} \tag{11}$$

This form makes it possible to evaluate $\lambda(s)$ or $\Lambda(s)$ extremely efficiently. The compensator for the Hawkes process can be given as

$$\Lambda(t) = \lambda T - \frac{\alpha}{\beta} \sum_{i=1}^{N(T)} [1 - e^{-\beta(T-t_i)}], \tag{12}$$

where T is the last time index in the interval of interest and $N(T)$ is the total number of observed events over the interval $[0, T]$. The compensator gives the expected number of events, denoted $\mathbb{E}[N(T)]$, over the interval of interest $[0, T]$.

Equation (7), giving the log-likelihood of the point process, can be extended to the Hawkes process with exponential decay. The log-likelihood for an exponentially decaying Hawkes process is given below:

$$\ell = \sum_{i=1}^{N(T)} \ln \left[\lambda + \alpha \sum_{j=1}^{i-1} e^{-\beta(t_i-t_j)} \right] - \lambda T - \frac{\alpha}{\beta} \sum_{i=1}^{N(T)} \left[1 - e^{-\beta(T-t_i)} \right]. \tag{13}$$

There is no analytic maximiser for the above equation, but due to the similar structure of the inner summations, the expression can be simplified. Let

$$A(i) = \sum_{j=1}^{i-1} e^{-\beta(t_i-t_j)},$$

such that $A(i)$ simplifies to

$$A(i) = e^{-\beta(t_i - t_{i-1})}(1 + A(i-1)),$$

With the base case of A(1) = 0, the log-likelihood of Hawkes with exponential decay can be rewritten as

$$\ell = \sum_{i=1}^{N(T)} \ln(\lambda + \alpha A(i)) - \lambda T - \frac{\alpha}{\beta} \sum_{i=1}^{N(T)} \left[1 - e^{-\beta(T - t_i)}\right]. \tag{14}$$

This form of the log-likelihood is more computationally efficient to evaluate.

3.4 Marked Hawkes process

The marked Hawkes process is an extension of the Hawkes which takes into account additional data about each arrival in the system, beyond the time of the arrival. This additional information can be incorporated in the form of marks, which influence the conditional intensity of the process. The following material, theorems, proofs and derivations in this section are largely based on those of Lapham (2014), who provides a detailed exploration of the marked Hawkes process.

The general conditional intensity of the marked Hawkes process can be given in a similar form to equation (9), as

$$\lambda(t|\mathcal{H}(t)) = \lambda + \sum_{i:t_i<t} \tau(t - t_i, m_i), \tag{15}$$

where $\tau(s, m) \geq 0$ for some $s \geq 0$, $m \geq 0$ and zero otherwise. The function $\tau(t, m)$ gives the effect of an arrival in the system and can be specified depending on the required effect of an arrival. It is common for

$$\tau(t, m) = g(m)\mu(t),$$

where $\mu(t)$ is the response function as given earlier in the section detailing the Hawkes process. The function $g(m)$ is referred to as the impact function and specifies the effect of the marks associated with events on the conditional intensity function. Exponential impact functions specified as $g(m) = e^{\delta m}$ are widely used and are chosen for this analysis.

The conditional intensity of a marked Hawkes process with an exponential decay function and an exponential impact function can be shown to be

$$\lambda(t|\mathcal{H}(t)) = \lambda + \alpha \sum_{i:t_i<t} \exp(\delta m_i - \beta(t - t_i)), \tag{16}$$

where $\alpha, \delta \geq 0$ and $\lambda, \beta > 0$ and the effects of arrivals before time point 0 are ignored. similarly to the Hawkes process, the α parameter gives the instantaneous excitation effect that an event has on the system, while β gives the rate of the decay of the function to the baseline intensity, λ. The δ parameter is involved in the impact function of the marks, and specifies the contribution of the given mark to the instantaneous increase in intensity induced by the occurrence of an event. The expected number of events over the interval $[0, T]$ can be given by the marked compensator,

$$\Lambda(t) = \lambda T - \frac{\alpha}{\beta} \sum_{i=1}^{N(T)} e^{\delta m_i}[1 - e^{-\beta(T-t_i)}], \tag{17}$$

where T is the last time index in the interval of interest and $N(T)$ is the total number of observed events over the interval $[0, T]$. The compensator gives the expected number of events, denoted $\mathbb{E}[N(T)]$, over the interval of interest $[0, T]$.

If the data consists of observations of points $(t_1, m_1), \ldots, (t_{N(T)}, m_{N(T)})$ from some finite interval $[0, T]$ as seen in Figure 1, then the log-likelihood of the marked Hawkes process can be given as

$$\ell = \sum_{i=1}^{N(T)} \ln \left[\lambda + \alpha \sum_{j=1}^{i-1} e^{\delta m_j - \beta(t_i - t_j)} \right] - \lambda T - \frac{\alpha}{\beta} \sum_{i=1}^{N(T)} e^{\delta m_i}[1 - e^{-\beta(T-t_i)}] + N(T) \ln(\gamma) - \gamma \sum_{i=1}^{N(T)} m_i, \tag{18}$$

where the marks are exponentially distributed with rate parameter γ, which is equal to the inverse of the mean of the marks. The marked Hawkes process with an exponential decay term and an exponential impact function has a Markov property, making the likelihood possible to be evaluated by the recursive function

$$A(i) = \lambda + e^{-\beta(t_i - t_{i-1})}(A(i-1) - \lambda) + \alpha e^{-\beta(t_i - t_{i-1})} e^{\delta m_{i-1}}, \tag{19}$$

where $A(i) = \lambda(t_i | \mathcal{H}(t_i))$ and $A(1) = \lambda$.

3.5 Moments of Hawkes processes

The expectation of the conditional intensity and counting process for the Hawkes process can be found via the derivation of the first moment of the respective functions. These can be compared to the observed functions. The first moment of the Hawkes intensity with exponential decay, given the initial intensity $\lambda(0)$ at the start of the interval, can be shown to be

$$m_1(t; \lambda_0) := \mathbb{E}[\lambda(t) | \lambda(0)] = \frac{\lambda\beta}{k} + (\lambda(0) - \frac{\lambda\beta}{k}) e^{-kt}, \tag{20}$$

where $k = \beta - \alpha$ (Laub, Lee, and Taimre, 2021, pg 63-64). Using the first moment of the conditional intensity given $\lambda(0)$, the expectation of the counting process, $N(t)$, conditional on $N(0) = 0$ and $\lambda(0)$ is given as

$$w_1(t; \lambda_0) := \mathbb{E}[N(t)|\lambda(0)] = m_1 t + (\lambda(0) - m_1)\frac{1}{k}(1 - e^{-kt}), \tag{21}$$

where $k = \beta - \alpha$.

Similarly, the expectations of the conditional intensity and counting process can be found for the marked Hawkes process (Cui, Hawkes and Yi, 2020, pg 104). The first moment of the marked Hawkes conditional intensity can be derived from the intensity found in equation (16), and is given as follows:

$$m_1(t; \lambda_0) := \mathbb{E}[\lambda(t)|\lambda(0)] = \frac{\lambda\beta}{k} + (\lambda(0) - \frac{\lambda\beta}{k})e^{-kt}, \tag{22}$$

where

$$\begin{aligned} k &= \beta - \mathbb{E}[\alpha e^{\delta m}] \\ &= \beta - \alpha\mathbb{E}[e^{\delta m}] \\ &= \beta - \alpha\left(\frac{\gamma}{\gamma - \delta}\right). \end{aligned}$$

γ is the rate parameter of the exponential distribution of the marks, which is equal to the inverse of the mean of marks. Using the first moment of the conditional intensity given $\lambda(0)$, the expectation of the counting process, $N(t)$, conditional on $N(0) = 0$ and $\lambda(0)$ for the marked Hawkes counting process is given as

$$w_1(t; \lambda_0) := \mathbb{E}[N(t)|\lambda(0)] = m_1 t + (\lambda(0) - m_1)\frac{1}{k}(1 - e^{-kt}), \tag{23}$$

where k is defined as above for equation 2 and m_1 relates to the first moment of the marked Hawkes conditional intensity.

3.6 Model fitting and diagnostics

The Hawkes and marked Hawkes models in this paper were built with the formulation seen in (10) and (16) respectively. The model parameters for all models, λ, α, β and δ, were found via maximum likelihood estimation from the respective likelihoods found in equations (13) and (18). These optimisations were performed via the `optim()` function in `R` (R Core Team, 2022), using the L-BFGS-B method, a quasai-Newtonian method, allowing the specification

of upper and lower bounds for parameters. The likelihood surface was found to be sensitive to the starting values of parameters, so a large number of sets of starting parameter values were used. The optimal parameter set is chosen such that the log-likelihood is the largest (Lapham, 2014). This increases the chances of the optimisation algorithm converging to a global maximum.

For the Hawkes parameters, a sequence of length 15 was used for each parameter (λ, α, β) resulting in 3375 combinations being run. For the marked Hawkes parameters, a sequence of length 8 was used for each parameter (λ, α, β) and a sequence of length 7 for the marks parameter (δ) resulting in 3584 combinations being run. A 95% confidence interval for the optimal parameters can be obtained using the variance-covariance matrix.

As an initial comparison between the Hawkes and marked Hawkes models, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) will be used (Zucchini and MacDonald, 2009). Although AIC and BIC will not be used to only select either the Hawkes and marked Hawkes, they will be used to identify the trade-off between the complexity due to the additional parameter in the marked Hawkes model. The AIC and BIC are respectively given by:

$$\text{AIC} = -2\ell + 2q \quad \text{and} \quad \text{BIC} = -2\ell + q\log(N(T)), \tag{24}$$

where ℓ is the log-likelihood of the fitted model, q is the number of parameters of that model and $N(T)$ is the total number of observed events over the interval $[0, T]$.

Diagnostic quantile-quantile plots can be constructed, enabling comparison between the observed distribution of event inter-arrival times and the inter-arrival times simulated from the proposed models. This analysis uses a simulation method based on the thinning of Poisson processes.

Theorem 1: Poisson Process Thinning

Let $N = (N(t) : t \geq 0)$ be a Poisson process. For each arrival of N, assign it to process N_1 with a probability p and to process N_0 with a probability $(1-p)$. The new processes N_1 and N_0 are independent Poisson processes with rates $p\lambda$ and $(1-p)\lambda$ respectively.

This can be generalised to the inhomogeneous case: if points are assigned to N_1 with time-dependent probability $p(t)$, then N_1 is an inhomogeneous Poisson process with with rate function $\lambda(t) = \lambda p(t)$. This can be used to simulate data from a theorised inhomogeneous Poisson process, which can be used for model validation.

The algorithm for this simulation method is given by Algorithm (1).

The c value in Algorithm (1) is a tuning parameter, which specifies how often the $M(t|\mathcal{H})$ value is updated. Small values of c result in more values being accepted in the algorithm,

thereby reducing the computational cost of simulation, however, if c is too small, the computational cost of frequently updating $M(t|\mathcal{H})$ is high. For the purposes of this simulation, we set c at a value of 1.

Algorithm 1 Simulation by thinning algorithm

Set $t \leftarrow 0$, $i \leftarrow 0$, $\mathcal{H} \leftarrow \emptyset$
while $t < T$ **do**
  Calculate $M(t|\mathcal{H}) = \lambda(t + |\mathcal{H})$ and $L(t|\mathcal{H}) = c\lambda(t + |\mathcal{H})$ for a chosen c;
  Simulate an exponential r.v R with a mean $\frac{1}{M(t|\mathcal{H})}$ **if** $R > L(t|\mathcal{H})$ **then**
    Set $t \leftarrow t + L(t|\mathcal{H})$
  else
    Simulate a uniform (0, 1) r.v. U;
    if $U > \frac{\lambda(t+R|\mathcal{H})}{M(t|\mathcal{H})}$ **then**
      Set $t + R$
    else
      Set $i \leftarrow i + 1$, $t \leftarrow t + R$ and $t_i = t$;
      For marked Hawkes : Simulate m_i exponential with rate γ;
      F(.);
      Set $\mathcal{H} \leftarrow \mathcal{H} \cup (t_i, m_i)$
  return $(t_1, m_1), (t_2, m_2), \ldots, (t_i, m_i)$

Further model diagnostics can be performed by making use of the random time change theorem, as given below. This is a goodness-of-fit test for the temporal component of a Hawkes process and involves rescaling the observed points of a Hawkes process via the compensator, such that the transformed time points are a unit-rate Poisson process.

Theorem 2: Random Time Change Theorem

Let $t_1, t_2, \ldots, t_k$ over the interval $[0, T]$ be the realisation of some point process with conditional intensity function $\lambda(t)$. If $\lambda(t)$ is positive over $[0, T]$ and $\Lambda(t) < \infty$, then the transformed set of points, denoted $\Lambda(t_1), \Lambda(t_2), \ldots, \Lambda(t_k)$, form a Poisson process with unit rate.

Theorem (2) was extended by Daley and Vere-Jones (2003) as follows.

Theorem 3: Residual Analysis

The compensator-transformed sequence $\Lambda(t_1), \Lambda(t_2), \ldots, \Lambda(t_k)$, denoted $t_1^*, t_2^*, \ldots, t_k^*$ and known as the 'residual process', whose counting process is $N^*(t)$, is a realisation of a unit-rate Poisson process if and only if the original set of events, $t_1, t_2, \ldots, t_k$, is from the point process defined by the compensator, $\Lambda(t)$.

Using theorem (2) and (3), observations of event times can be transformed such that the

resulting counting process can be compared to the expected counting process under a unit-rate Poisson distribution.

The distribution of the inter-arrival times of the residual process, $\epsilon_{i+1} = t^*_{i+1} - t^*_i$, can also be tested as a goodness-of-fit metric. The residual inter-arrival times should be exponentially distributed with a unit mean if the model is a suitable fit for the data.

A test described by Ogata (1988) and credited to Berman (1983), involves transforming the residual inter-arrival times as follows:

$$U_k = 1 - \exp(-\epsilon_i),$$

where the U_k variables should be i.i.d. distributed uniform variables on the interval (0,1). A test of the independence between consecutive intervals can be performed by plotting the points (U_k, U_{k+1}), which should be a random scatter in the $(0,1) \times (0,1)$ space if the model is a good fit for the data.

4 Data description

4.1 Site and measurements

The dataset is from the South African Environmental Observation Network (SAEON, 2022) and contains information from a Savanna site near Kimberly in the Northern Cape province, South Africa (latitude -28.890600, longitude 24.861117, and altitude 1183m). The site consists of open savanna vegetation of grass and trees, with a maximum height of 0.7m and 8m, respectively. Various instruments were used to measure the meteorological and flux data variables, while the Eddy Flux Data for KIMTRI (Kimberley area, Northern Cape and Western Free State) were recorded from an IRGASON instrument at a height of 3.5m. These measured fluxes are vertical turbulent fluxes within atmospheric boundary layers.

The Eddy Flux observations are a direct method to observe the exchange between the ecosystem and atmosphere in terms of gas, energy, and momentum (Liang and Wang, 2019). These measurements are often used to understand how weather conditions arise to aid in the understanding of weather measurements at particular points.

Campbell Scientific Inc. designed the single-sensor IRGASON which is an integrated open-path infrared gas analyzer (IRGA) and sonic (SON) anemometer (Campbell Scientific, 2022a). The IRGASON is specifically designed to measure the eddy-covariance carbon and water flux measurements.

Figure 2: IRGASON instrument designed specifically for eddy-covariance flux measurements (Campbell Scientific, 2022b).

The CO_2 raw variable was recorded by the IRGASON and was sampled at 20Hz before being converted to 30-minute fluxes. This carbon dioxide mole fraction in wet air has the unit mmol per mol and can be viewed as the carbon dioxide density. SAEON is interested in CO_2 monitoring because of its relevance to climate change as it is a major greenhouse gas.

4.2 Data management

The dataset contains 68 variables and 34,520 half-hourly observations recorded between 14/1/2020 and 2/1/2022. The data received from SAEON has undergone automated extended quality control. The automated quality control procedures include; screening the data for outliers, identifying and correcting periods of instrument drift from baseline, identifying and correcting periods of repeated values and correcting or removing any negative values.

Missing values do occur when calibration and maintenance periods were performed, but will not be removed from the dataset as the missing values should generally not be removed in time series data. Moreover, the timing of the extreme CO_2 events is fundamentally important when modelling the Hawkes and marked Hawkes processes, and as such removing the missing values will impact the intensity estimate in the processes. When modelling the conditional intensity, the missing values will be treated as no arrival occurring at that particular time.

5 Exploratory data analysis

CO_2 exploration

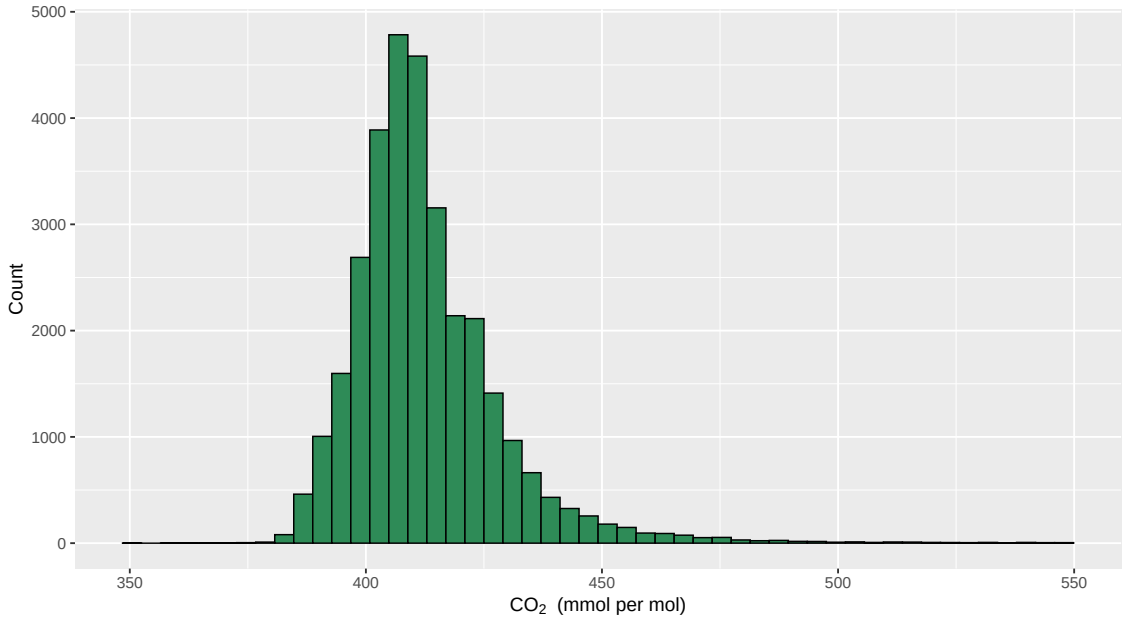

Figure 3: Skewed distribution of the CO_2 variable with a long right tail.

The data is narrowly peaked with a long tail to the right (Figure 3). This narrow peak is relative to the large range of observed CO_2 values plotted along the x-axis, with the minimum and maximum observed CO_2 values given by the table below to be 351.904 and 549.3, respectively. Looking at the five-number summary of the data given in Table 1, the range of values in the first 50% of the CO_2 distribution is approximately 58 mmol/mol. If this is compared to the range of the second half of the distribution, which is approximately 139 mmol/mol, it is clear that the right-hand side of the distribution is skewed, covering a larger range of values with the same number of observed data points. This long tail on the right indicates the presence of extreme values on the upper end of the distribution. Additionally, the kurtosis is fairly large, indicating the distribution is leptokurtic, which may be attributed to the extreme values in the tail.

Table 1: Numerical summary statistics of CO_2 variable, where Min. is the minimum, Q_1 is lower quartile, Q_2 is upper quartile, Max. is the maximum, St. Dev. is the standard deviation and Kurt. is kurtosis.

	Statistic							
Variable	Min.	Q_1.	Median	Q_3.	Max.	Mean	St. Dev.	Kurt.
CO_2	351.904	403.015	409.863	419.528	549.300	412.665	15.891	11.769

In addition, the fluctuation of climatic CO_2 levels between wet and dry seasons is well documented in literature, so there may be higher extremes in the wet season than in the dry (Kim, Verma, and Clement, 1992; Hovenden, Newton and Wills, 2014; Räsänen et al, 2017). Therefore the extreme values that lie within the wet and dry seasons need to be identified.

In Kimberly, the wet season lasts a total of 5.9 months, from October 19 to April 15 while the dry season lasts 6.1 months, from April 15 to October 19 (Weather Spark, n.d.). A wet day is defined as a day with at least 0.04 inches of liquid precipitation. Therefore the wet season consists of days that have a greater than 15% chance of being considered a wet day. On average February is the month that has the wettest days while July has the fewest wet days.

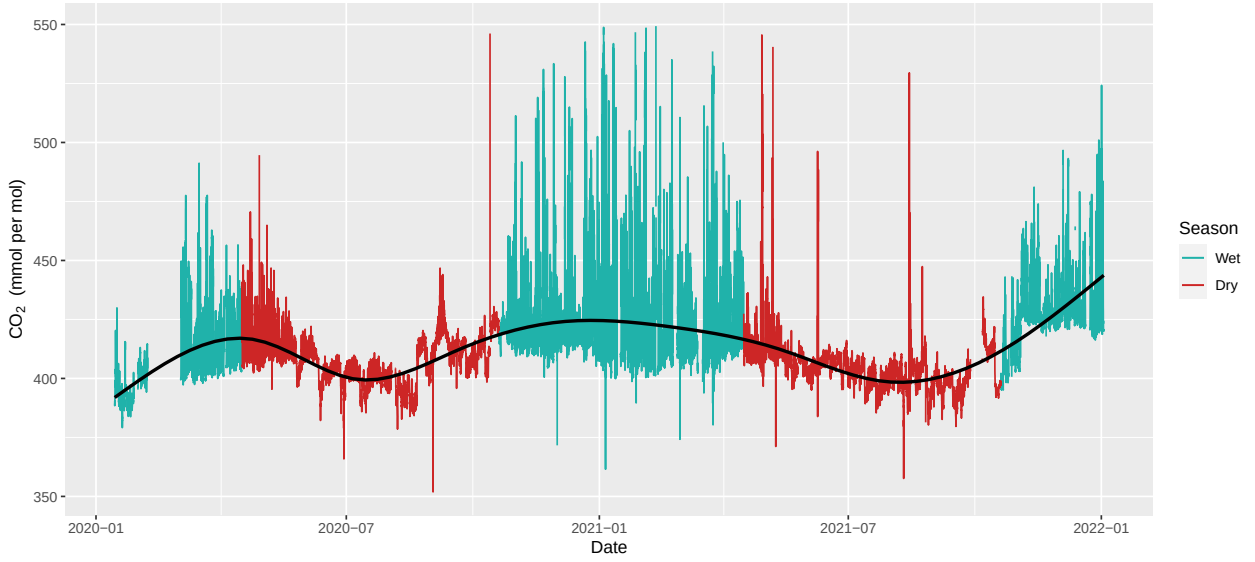

Figure 4: CO_2 level over 2 years separated by the wet and dry seasons, with a smoothed spline from a generalised additive model (black line).

There are two complete dry seasons and only one complete wet season available in this dataset. Seasons are considered complete when they contain observations spanning the entire seasonal interval. The smooth spline suggests that the average CO_2 changes over the course of a year and that the CO_2 levels are different in the wet and dry seasons. Dry seasons 1 and 2 have respective CO_2 averages of 407.997 and 404.142, which are lower than all of the CO_2 averages for the complete and incomplete wet seasons, even though the wet season 1 average of 408.68 is only marginally larger than those of the dry season (Table 2). This low average may be due to missing values or that this particular wet season experienced less rainfall than normally observed. It is also worth noting that this wet season is incomplete and may have experienced heavier rainfall previously, which could have resulted in a higher average CO_2 value. The wet season has more common extreme CO_2 values, which indicates environmental clustering. As the level varies throughout the year, the frequency of extremes present are be season dependent.

Table 2: Season interval start and end dates with the corresponding number of observations (excluding missing value) with the average CO_2 for that season.

Season	**Interval start**	**Interval end**	**Number of observations**	**Average CO_2**
Wet 1	1/14/2020	04/15/2020	3027	408.68
Dry 1	04/16/2020	10/19/2020	8842	407.997
Wet 2	10/20/2020	04/15/2021	7798	422.385
Dry 2	04/16/2021	10/19/2021	8502	404.142
Wet 3	10/20/2021	1/2/2022	3305	422.385

As there is seasonality present in CO_2 observations that need to be accounted for, the data will be split into the three complete seasons, with a separate model for each. This seasonal splitting is used in a similar manner to that of Danquah, Sigauke and Kyei (2020). An aim of this analysis is to understand how the extreme values behave in each season, thereby how they change over time, so the seasonality is accounted for by splitting the data. The three seasons that will be modelled are dry 1, wet and dry 2. Wet season 2 will now be referred to as 'Wet season'.

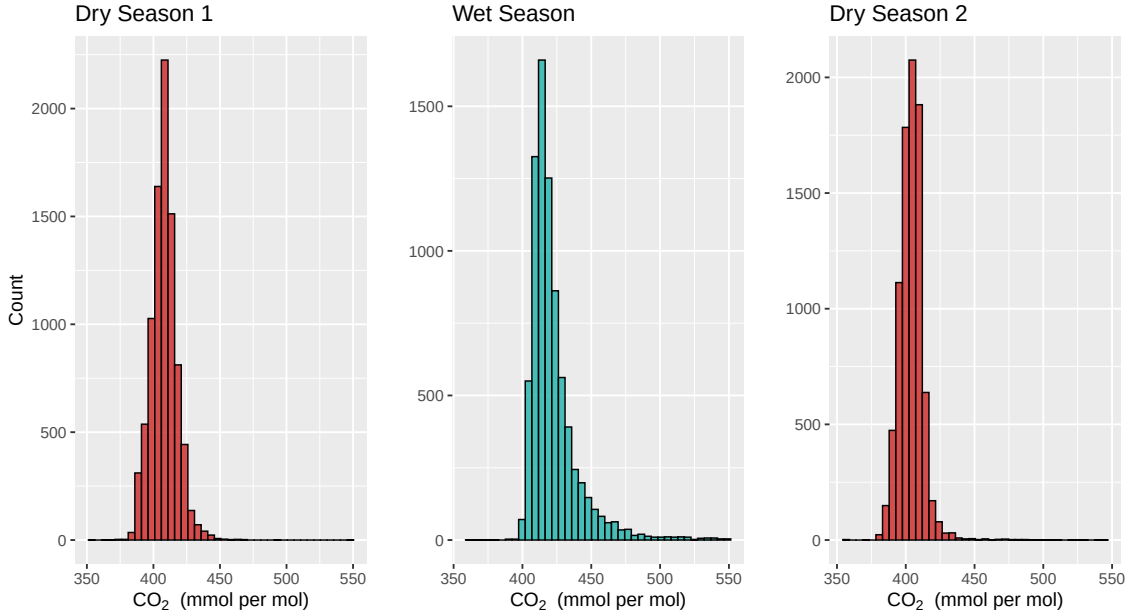

Figure 5: Distributions of the CO_2 variable, by the three complete seasons; dry 1, wet and dry 2.

All three distributions are narrowly peaked relative to their respective range of CO_2 values, with the maximum CO_2 value for the first dry, wet and second dry seasons being 546.168, 549.300 and 545.56 respectively (Table 3). For dry season 1, the range of values covered by the first 50% of the distribution is approximately 62 mmol/mol, while the range covered by the second 50% of the distribution is about 139 mmol/mol. This indicates that the right-hand

side of the distribution is skewed, covering a larger range of values with the same number of observed data points. This long tail on the right implies the presence of extreme values on the upper end of the distribution, which are the events of interest to be modelled. The same relationship can be observed in both the wet and second dry season CO_2 values, with the range covered by the first half of each seasonal CO_2 distribution being approximately 56 mmol/mol and 46 mmol/mol respectively, while the range of the second half of each distribution is approximately 132 mmol/mol and 142 mmol/mol.

Table 3: Numerical summary statistics of CO_2 for the three complete seasons; dry 1, wet and dry 2.

	Statistic							
Season	Min.	1st Qu.	Median	3rd Qu.	Max.	Mean	Std. Dev.	Kurt.
Dry 1	351.904	401.918	407.977	413.380	546.168	407.997	9.832	9.596
Wet	361.511	411.670	417.370	426.892	549.300	422.385	18.080	13.237
Dry 2	357.615	398.590	404.054	409.041	545.562	404.142	9.454	36.336

The above is supported by the kurtosis values for each season (Table 3). All three seasons have large kurtosis values of 9.596, 13.237 and 36.336 for the first dry, wet and second dry seasons respectively. These values are much larger than the expected normal distribution kurtosis of 3, once again indicating the distributions are leptokurtic. This is expected as these are subsets of the whole leptokurtic dataset. The three seasons have a similar range with varying quartiles and interquartile ranges, which is explored further using figure 6.

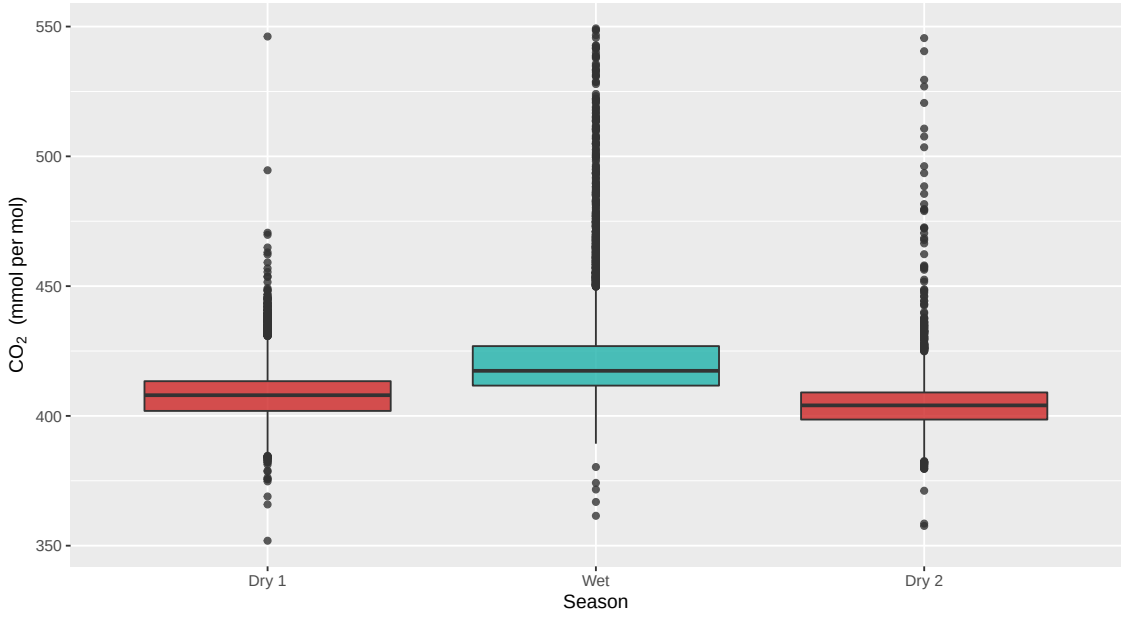

Figure 6: Boxplot showing the spread of the CO_2 variable, for the three complete seasons; dry 1, wet and dry 2.

Table 3 and Figure 6 show that the mean CO_2 value of 422.385 is larger in the wet season than in either of the dry seasons where dry season 1 has a mean of 407.997 and dry season 2 has a mean of 404.142. For all of the five number summary statistics, the wet season has larger values than either of the dry seasons. This once again indicates there is seasonality present in the data. The interquartile range of dry season 1 is 11.462 and dry season 2 has a similar value of 10.451 but dry season 1 appears to have fewer outliers than dry season 2. Dry season 2's outliers cover a larger range than those of dry season 1. Now a threshold for which exceedances are considered extreme must be determined to get the half-hourly CO_2 recordings into suitable point process form.

Before exploring the threshold, considerations regarding the data resolution were made. Three options were considered; taking a daily maximum, a daily mean or finding a day and night average. The daily maximum approach comes from extreme value theory whereby a block-maximum approach is used to find a per-period maximum; but this results in reduced dependence amounts events as well as loss of information contained in the data (Gebbie, 2022). If the daily mean or maximum is taken, the overall dataset size is reduced from 34520 observations to 720 observations. For the day and night option, the overall dataset would be 1440 observations. This will result in a smaller dataset, but a significant loss of information regarding the extreme readings.

Moreover, an issue arises as the natural clustering of events is lost as well as other information relevant to the system. As the dependence between events is of interest in the modelling process, these methods will not be used and the resolution will be kept as it is.

Another consideration is that the natural way to define an event as extreme would be to choose the initial time at which the CO_2 level first exceeded the threshold and then record its duration. This might result in the events losing their clustering, thereby losing their dependence, and spanning more than a one-time point, which is undesired and cannot be modelled by the simple Hawkes or marked Hawkes processes.

Importantly, the CO_2 exceedances are most likely not triggering new exceedances. Therefore the use of a self-exciting Hawkes process may not appear to be an intuitive choice of model. However, a self-exciting Hawkes process may be able to capture underlying environmental conditions influencing the frequency of CO_2 events over time. In order to further attempt to capture an underlying environmental condition, CO_2 exceedance magnitude marks will be included in the Hawkes formulation.

Threshold exploration and determination

To determine a suitable threshold, two techniques are utilized. The first is a graphical diagnosis using a mean residual life plot with the 90th percentile plotted as a comparison. This plot gives the numbers of exceedances for a particular threshold as well as the mean excess. If the assumption is correct that the GPD is a valid model for the exceedances, then the plot should be approximately linear above a threshold u (Bommier, 2014; Gebbie, 2022).

Although the aim is not to model the distribution but rather the intensity of extreme events, this diagnosis can still be used as it identifies and captures the right-tail values considered extreme. The second technique for determining the threshold is using the rule of thumb.

In order to determine the seasonally dependent threshold, all seasonal observations were used such that the wet season comprised of both non-complete and complete wet seasons, and the dry comprised of both complete dry seasons. This approach was used in an effort to produce a more broadly representative seasonal threshold, less influenced by the observations in the particular season in question.

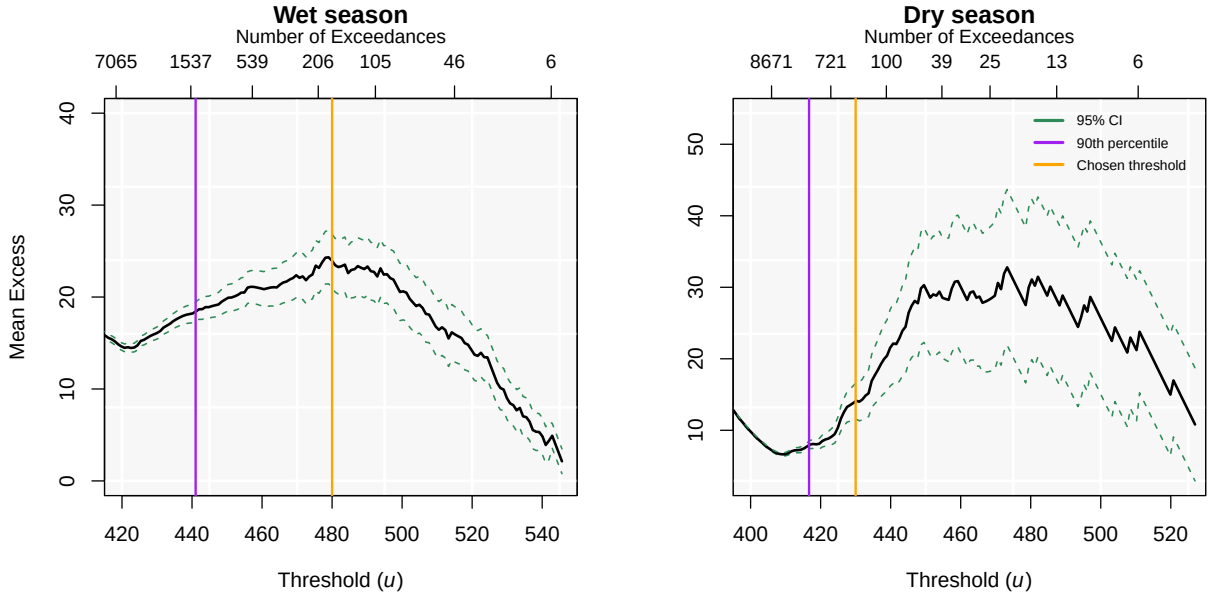

Figure 7: The mean residual life plot of the CO_2 variable for the wet (left) and dry season (right), with the empirical mean (solid black jagged line), 95% central limit theorem-based confidence intervals. For comparison, the 90th percentile and the chosen threshold which is used to extract the point process realisation are given.

The mean excess for a given threshold is the average magnitude of the exceedances beyond that particular threshold. It is expected that the mean residual life plot becomes unstable towards the right due to few data points observed and therefore fewer excesses. This would result in a wide volatile 95% confidence interval, as seen in the dry season (Figure 7). However, the interval gets narrower for the wet season. This narrowing interval occurs because as the threshold increases, the CO_2 extreme values in this data set happen to have a similar magnitude above that threshold (with a smaller range), and as a result, the standard error decreases. Subsequently so does the confidence interval, due to all the values getting closer to the mean. As there are few data points towards the right of both the plots in Figure 7, it is highly unstable and a threshold that results in a few number of exceedances will not be considered.

When deciding on the threshold value, both the mean excess and number of exceedances must be taken into account. For the wet season, the threshold value above which the exceedances are approximately linear is around 480. Although a threshold value lower than 480 is also approximately linear, it would result in a larger number of exceedances with a lower mean excess that is not representative of the right-tail extremes that can be considered truly climatically extreme events. The 90th percentile suggests a low threshold value of approximately 441 which results in a large number of extreme events, hence it is not appropriate.

For the dry season, the threshold value is around 430 where the mean excess is approximately linear after this point until it gets unstable due to the low number of CO_2 exceedances. Once again the 90th percentile suggests a low threshold of approximately 418 which results in too many events being considered extreme. It is important to note that the number of CO_2 events considered extreme for the dry seasons will be split between dry seasons 1 and 2. As this plot can be difficult to interpret, the rule of thumb will be used in conjunction with these values.

Table 4: Threshold calculated by the rule of thumb for each season that has n observations, with the corresponding number of exceedances, where η is the number of events above the threshold.

Season	**Rule**	**Threshold**	**Number of exceedances**
Wet	$\eta = \sqrt{n}$	488.427	119
	$\eta = n^{\frac{2}{3}}/\log(\log(n))$	472.236	259
Dry	$\eta = \sqrt{n}$	436.058	132
	$\eta = n^{\frac{2}{3}}/\log(\log(n))$	428.326	294

Firstly, the $\eta = \sqrt{n}$ rules give a higher threshold for both the wet and dry seasons than those previously identified from the mean residual life plot of 480 and 430 respectively. A higher threshold naturally results in a smaller number of exceedances and therefore a smaller number of observations which are considered extreme. On the other hand, the $\eta = \sqrt{n}$ rule for the dry season suggests a threshold value of 428.326 which is similar to that identified in figure 7 which was 430. For the wet season, the threshold value of 480 identified earlier is in between the two threshold values of 436.058 and 428.326 given by the rule of thumb.

Therefore, combining the results from the mean residual life plot and the rule of thumb, the chosen threshold for the wet season is 480 mmol/mol and 430 mmol/mol for the dry season. This results in dry season 2 having 77 fewer events than dry season 1 (Table 5) such that the first dry season contains more values that are considered dry extremes. The value of the CO_2 exceedances above the threshold will be used as a mark.

Table 5: Chosen threshold for each season with the corresponding number of extreme events given above that threshold.

Season	Chosen threshold	Number of extreme events
Dry 1	430	164
Wet	480	140
Dry 2	430	87

The data does appear to show clustering of the extreme events above the selected threshold due to the points being temporally close together (Figure 8). These tight temporal clusters are especially evident with both of the dry seasons, as the initial observations are tightly clustered and then a large amount of time passes before the next cluster occurs. It is worth noting that this clustering at the beginning of the season could be due to lagged effects from rainfall in the wet season.

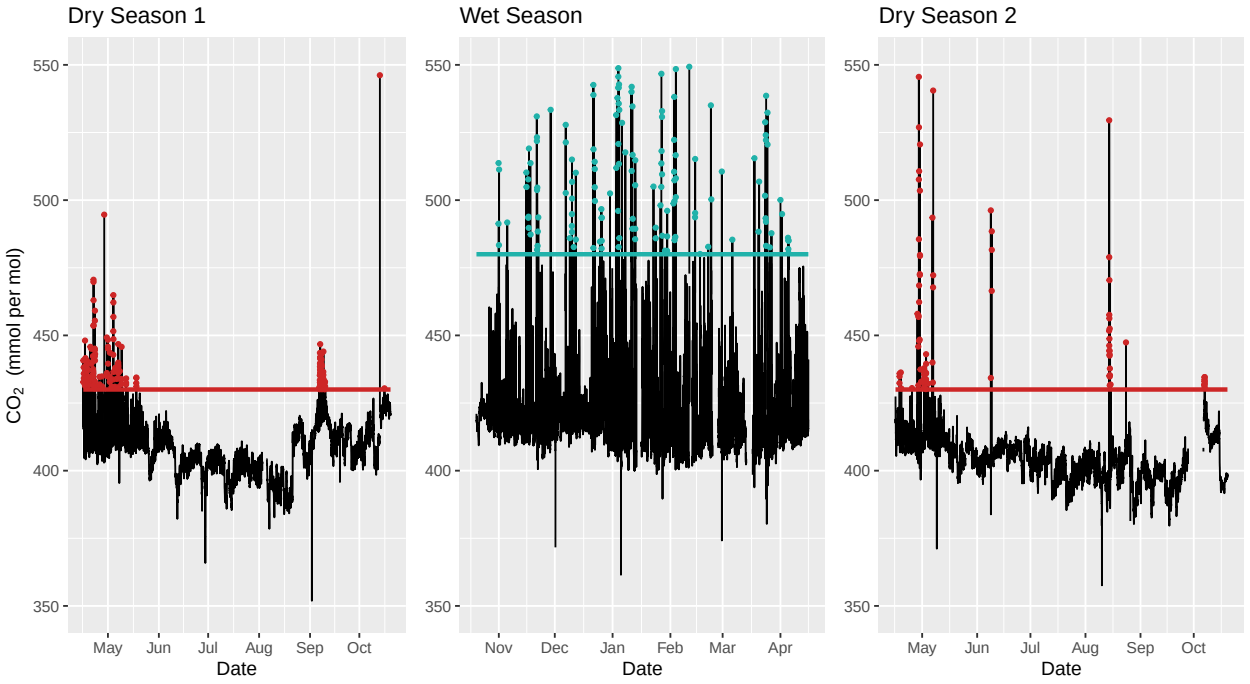

Figure 8: Extreme events exceeding the threshold of 480 for the wet season and 420 for both of the dry seasons. Points indicate an extreme event and the horizontal line is the chosen threshold. Dry season 1 has 164 extreme events, the wet season has 140 extreme events and dry season 2 has 87 extreme events.

From here onwards, the three seasonal datasets only containing events that are considered extreme will be referred to as 'Dry extreme 1', 'Wet extreme' and 'Dry extreme 2'. Hence the extreme events within these datasets will be referred to as events. An important part of modelling point processes is the inter-arrival times between these events.

Inter-arrival Times

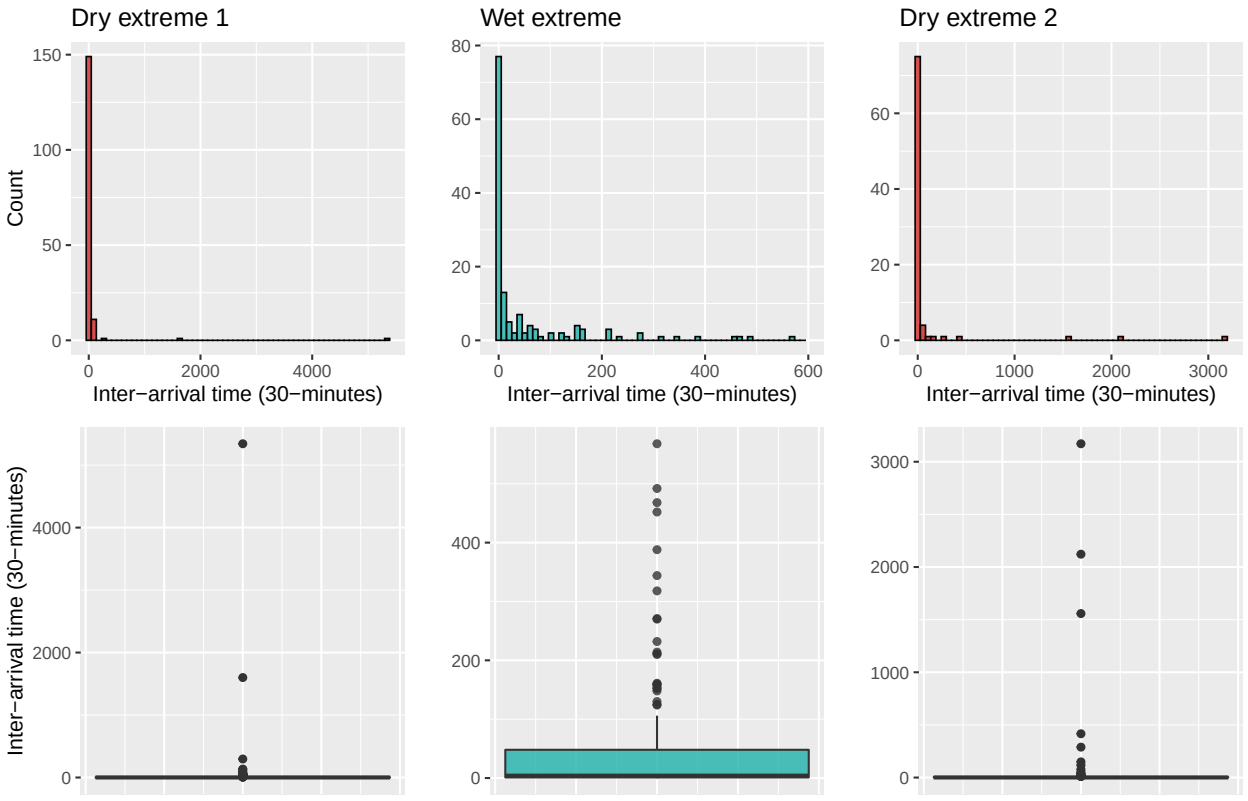

Figure 9: Histogram and box plots of the distribution of the inter-arrival times for the wet and dry extremes, for comparison.

The inter-arrival times between the CO_2 events differ greatly between the wet and dry extremes, where 1 unit is equal to a 30-minute interval. The first dry extreme's inter-arrival time has the largest range with a maximum value of 5342, although this is a large outlier in terms of the inter-arrival times (Table 6). This outlier may affect the intensity estimation. For both the dry extremes 1 and 2, 75% of the inter-arrival times were 1 unit apart (30 minutes) leading to highly right-skewed data. For the wet season, only 50% of the inter-arrival times were 1 unit apart (30 minutes). The wet extreme inter-arrival times are also skewed right but not as severely as the dry extreme inter-arrival times. This indicates that there is more tight clustering with larger inter-arrival times in both of the dry extremes, whereas wet has more clusters of smaller size with less time in between them as seen in Figure 8.

Table 6: Numerical summary statistics of the inter-arrival times for the wet and dry extremes.

	Statistic						
Extreme	Min.	1st Qu.	Median	3rd Qu.	Max.	Mean	Std. Dev.
Dry 1	1	1	1	3	5342	53.963	436.029
Wet	1	1	4	48	568	53.698	107.294
Dry 2	1	1	1	4	3171	95.965	440.591

Moreover, the wet CO_2 extremes have a much smaller range of 567 units and a standard deviation of 107.294 than the dry CO_2 extremes. This may be due to the events being more regularly occurring than the dry events. As a marked Hawkes process will be used to model the intensity of the event in order to attempt to capture some underlying environmental condition, the marks distribution must be explored and an appropriate theoretical distribution identified.

Magnitude marks

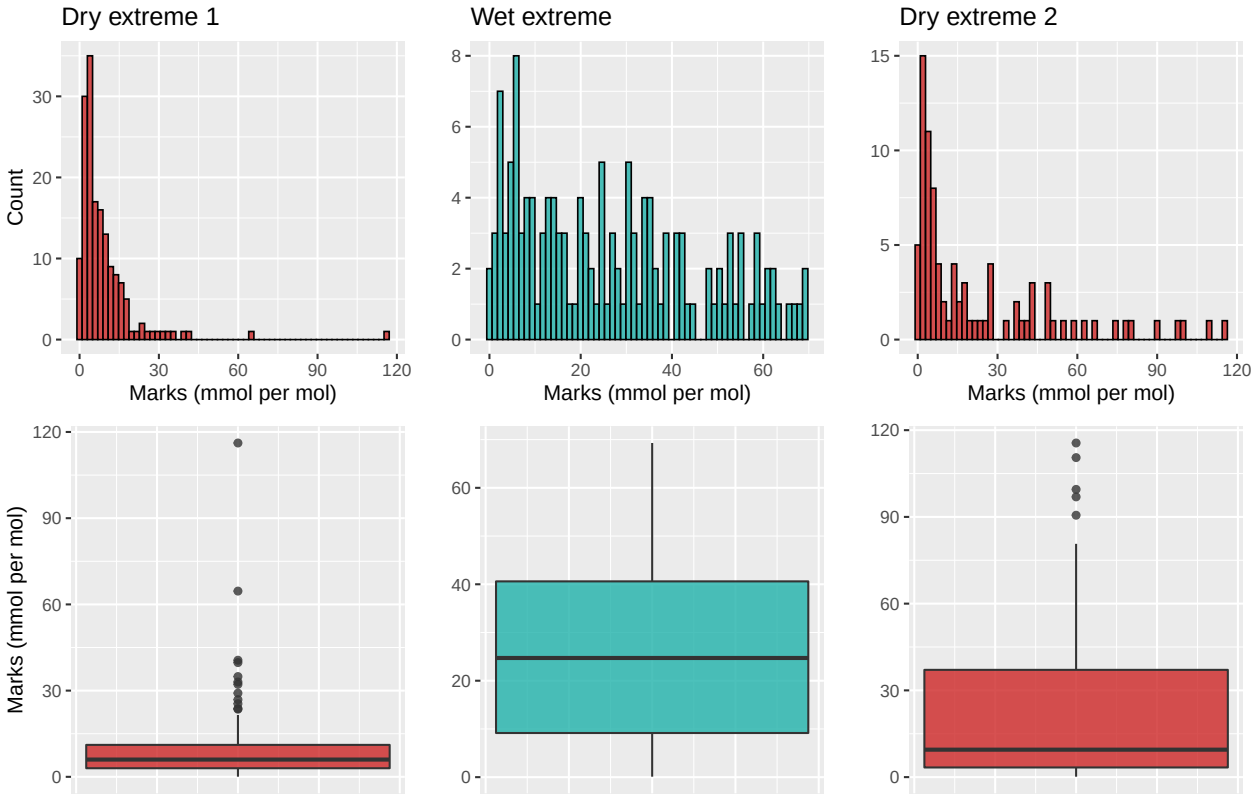

Figure 10: Histogram and box plots of the distribution of the marks for the wet and dry extremes, for comparison.

The dry extreme 1 marks have the most right-skewed distribution with 75% of the marks being below 11.174. Dry extreme 1 also has a large outlier, making the maximum 116.168 (Table 7). Both wet extreme and dry extreme 2 have more varied mark values demonstrated by their larger standard deviations of 19.9664 and 28.374 respectively. It is important to note that the dry extreme 1 marks are less spread and have a more homogeneous magnitude which can be seen by the smaller standard deviation around the mean than for wet and dry extreme 2 (Table 7).

Table 7: Numerical summary statistics of the CO_2 marks for the wet and dry extremes.

	Statistic						
Extreme	Min.	1st Qu.	Median	3rd Qu.	Max.	Mean	Std. Dev.
Dry 1	0.046	2.985	6.009	11.174	116.168	9.166	12.178
Wet	0.097	8.906	24.713	40.657	69.300	26.743	19.664
Dry 2	0.148	3.339	9.496	37.097	115.562	23.324	28.374

For both the dry extremes, the distribution appears to resemble an exponential distribution, while the wet extreme has no clear distribution. To determine a suitable distribution for the marks, an exponential distribution is simulated with a rate of μ^{-1} where μ is the mean of the extreme marks.

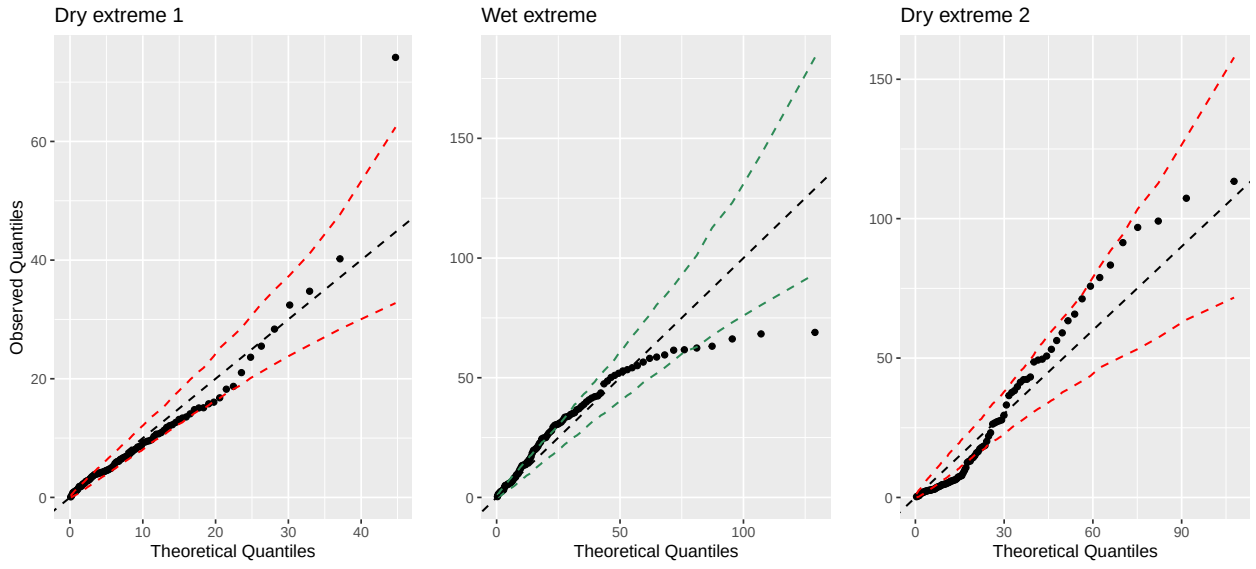

Figure 11: Quantile-quantile plot for the marks of each extreme dataset. Theoretical distribution simulated via the exponential distribution with a rate of μ^{-1} where μ is the mean of the extreme marks. The black dotted line is the Q-Q line, while the 95% simulated confidence intervals are given by the red dotted line for the dry extremes and by the green dotted line for the wet extremes.

For dry extreme 1, the majority of the points either fall within or just outside the 95% simulated confidence interval expected under the proposed exponential distribution, aside from the last point (figure 11). This last point is the maximum value of the marks, which previously has been identified as an outlier. The proposed exponential mark distribution does not appear to be an unreasonable fit for the data of dry extreme 1.

Similarly, for dry extreme 2, all the points either fall within the 95% simulated confidence interval or just outside. This suggests that an exponential distribution is an appropriate fit for the dry extreme 2 marks.

Lastly, for the wet extreme, the majority of the points either fall within the 95% simulated confidence interval or just outside, aside from the last 4 points. Although the observed quantiles deviate from the theoretical quantiles towards the tail of the distribution, the exponential distribution may still be a reasonable distribution as it appears suitable for both the dry extremes. Therefore an exponential distribution with rate μ^{-1} will be used, where μ is the mean of the extreme marks.

As the dry 1, wet and dry 2 extreme datasets have been thoroughly explored, they will be modelled using various point processes and thereafter the models fit will be checked and compared.

6 Dry extreme 1 application and results

6.1 Simple Poisson process

The intensity function for a simple point process stays constant at a value of λ, which can be calculated by dividing the total number of events by the complete time interval. In this way, the intensity function is unrelated to the current time or the history of events. The lambda test statistic for the first dry extremes is:

$$\lambda = 0.0185$$

This λ can be understood as the expected number of events at any given time in the interval is 0.0185, indicating that the occurrence of an event is very unlikely. The suitability of this model can be evaluated by assessing the underlying assumption of the model that the event inter-arrival times are exponentially distributed with a rate of $\frac{1}{\lambda}$. Given in the below-left plot is a quantile-quantile plot contrasting the observed inter-arrival times against simulated inter-arrival times from an exponential distribution with a rate of $\frac{1}{\lambda}$. This can be interpreted in conjunction with the plot below on the right. This gives the observed distribution of the inter-arrival times against a simulated distribution of inter-arrival times under the aforementioned assumptions of the Poisson model.

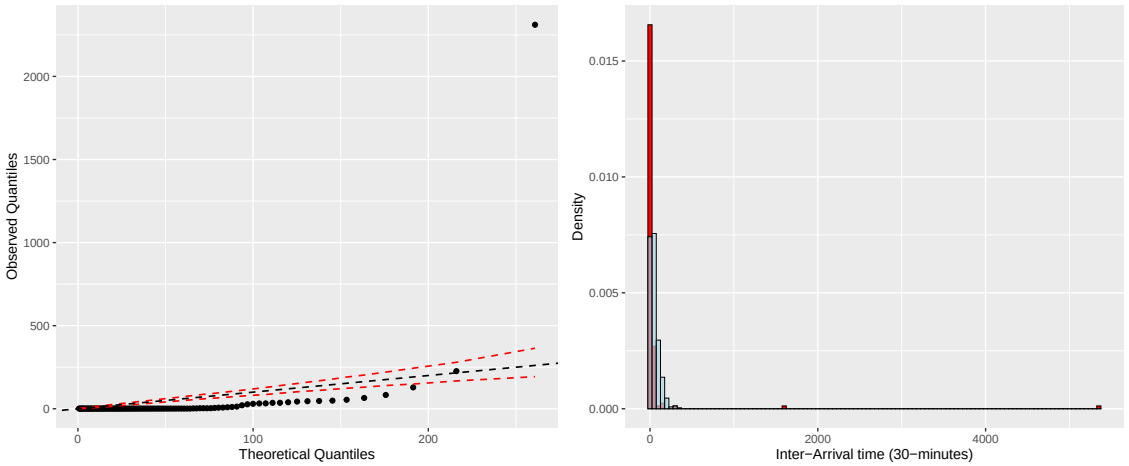

Figure 12: Quantile-quantile plot (left) of the observed inter-arrival times against a theoretical exponential distribution with the rate $\frac{1}{\lambda}$. The black dotted line is the Q-Q line, while the red dotted line gives a 95% simulated confidence interval. The right-hand plot gives the observed distribution of the inter-arrival times (red) against a simulated inter-arrival time sample of size 1000 from an exponential distribution with rate $\frac{1}{\lambda}$ (blue).

As seen by the above Q-Q plot, the vast majority of the plotted points fall below the simulated 95% confidence interval. This shows that the observed inter-arrival times are mostly smaller than what would be expected under a Poisson model, indicating the presence of clustering in the data which is not captured by the Poisson model. This observation is supported by the right-hand plot, which shows the distribution of the observed inter-arrival times against sampled inter-arrival times from a reference exponential distribution with a rate of $\frac{1}{\lambda}$. The observed distribution of inter-arrival times drops off sharply, indicating that the reference exponential distribution overestimates the inter-arrival times between events and fails to capture the clustering of extreme events that are observed in the data. A demonstration of the Poisson process model's failure to capture extremes can be done by using the reference distribution of inter-arrival times to simulate occurrences of events in time and compare the consistency, rate and spread of the simulated events against the observed events.

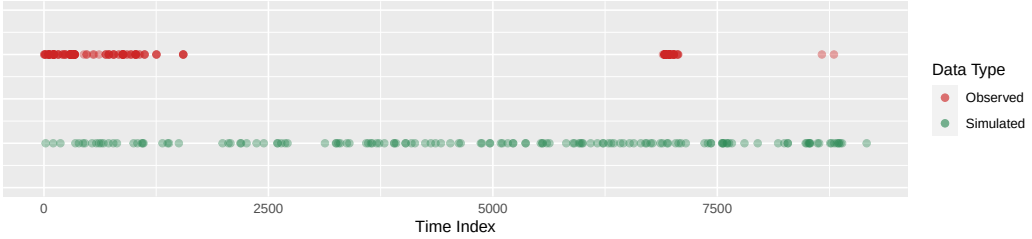

Figure 13: Dry extreme 1 observed events (red) with simulated events (green) over the same time interval. Simulated events via an exponential distribution with rate λ. There are 164 observed events and 147 simulated events.

The simulated Poisson process results in 147 events occurring over the interval of interest. This is an underestimate of 17 events. However, it is clear from the plot that the simulated distribution fails to capture the intense clustering observed in the observed data. The Hawkes process model can account for clustering in data by producing a conditional intensity which is dependent on the history of events.

6.2 Hawkes process

The Hawkes conditional intensity with exponential decay for the following parameters was calculated using the formulation of the Hawkes model in equation (10). The parameters were found by maximising the log-likelihood found in equation (14), which resulted in a log-likelihood value of -486.622.

Table 8: Optimised Hawkes parameters for dry extreme 1 with corresponding 95% confidence interval.

Parameter	**Value**	**95% confidence interval**
λ	0.0039	(0.0036, 0.0041)
α	0.323	(0.267, 0.379)
β	0.409	(0.351, 0.468)

The parameter λ (0.0039) specifies the baseline intensity of the process, which is the intensity towards which the process decays after the initial excitation effect of an event, which is given by α. The α parameter can be thought of as the instantaneous increase in the conditional intensity of 0.323 induced by each arrival on the system. The β parameter gives the rate of the exponential decay (0.409) with which the conditional intensity tends towards λ following the occurrence of an event. The 95% confidence interval associated with the λ parameter has a range of 0.0005, while the respective ranges for the 95% confidence intervals for the α and β parameters are 0.112 and 0.117. The difference in interval size can be attributed to the relative size difference between the λ parameter and the α and β parameters. The aforementioned parameters are given in the Hawkes conditional intensity in the below formula.

$$\lambda(t) = 0.0039 + \sum_{t_i<t} 0.323 \times e^{-0.409(t-t_i)}$$

The conditional intensity plot below is a subsection of the conditional intensity plotted over the interval. The conditional intensity over the whole time interval for the first dry season can be found in the Appendix (Figure 33). The intensity increases instantaneously upon the occurrence of an event and decays towards the baseline thereafter. The dotted orange line gives the expectation of the CO_2 intensity which is found by evaluating the first moment of the intensity function.

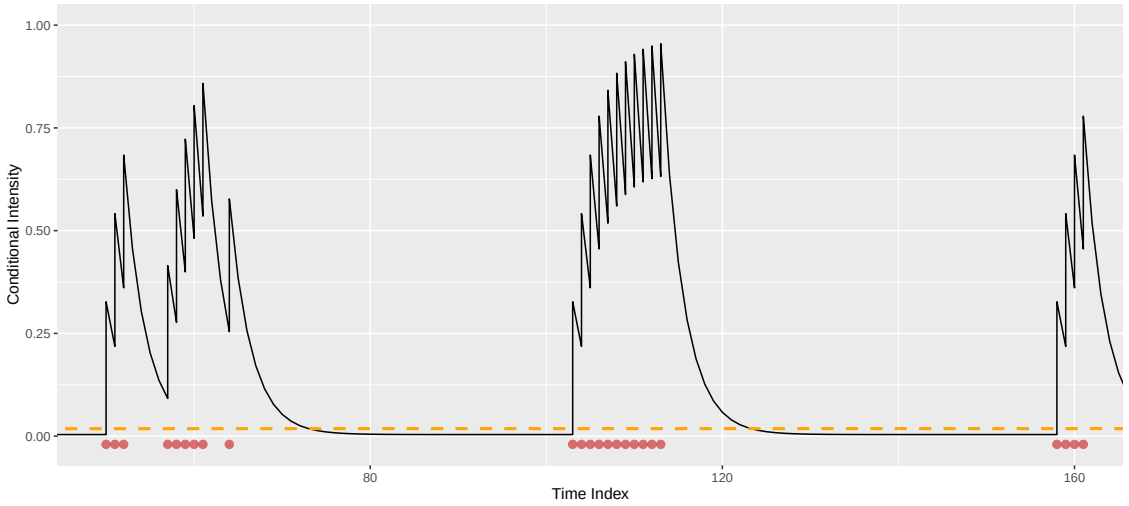

Figure 14: Subsection of dry 1 extreme Hawkes conditional intensity with parameters $\lambda = 0.0039$, $\alpha = 0.323$ and $\beta = 0.409$ given by the black line. Events are given by the red dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

6.3 Marked Hawkes process

The parameters for the marked Hawkes process model for the first dry season were found by maximising the log-likelihood in equation (19). The parameter values resulting the log-likelihood of -1013.961 are given in Table 9. The meaning of the parameters can be understood in the same way as previously described in the model formulation of the Hawkes model. The α parameter gives the instantaneous excitation effect of an arrival on the system, while the β parameter gives the rate of the decay of the conditional intensity towards the baseline intensity, λ. The added parameter, δ, gives the instantaneous impact effect that the magnitude of the CO_2 exceedances have on the conditional intensity. It is worth noting that this parameter for the first dry season was found to be approximately zero, indicating that the CO_2 magnitude mark does not appear to play a role in the conditional intensity. The other parameters are equal in value to those in the Hawkes process model for the same data. This is because, given that the CO_2 magnitude marks do not play a role, the model is equivalent to a Hawkes process model.

Table 9: Optimised marked Hawkes parameters for dry extreme 1 with corresponding 95% confidence interval.

Parameter	**Value**	**95% confidence interval**
λ	0.0039	(0.00378, 0.00395)
α	0.323	(0.283, 0.363)
β	0.409	(0.368, 0.451)
δ	1×10^{-5}	$(-3.73 \times 10^{-5}, 3.75 \times 10^{-5})$

The λ parameter has a much smaller chosen value and much narrower associated confidence interval than those of the α and β parameters (Table 9). The confidence interval for the λ parameter spans a range of 0.00017, while the intervals associated with the α and β parameter values are just below 0.1 each. This can be understood by considering the relative size of the parameter estimates, with the λ parameter being approximately a hundredth of the size of α and β. This indicates that the baseline intensity that the system decays to is very small. The result of an event on the system is an instantaneous increase in the intensity of 0.323, before decaying exponentially at a rate of 0.409 thereafter. It is worth noting that, although the λ, α and β parameters are equal to those of the Hawkes process model, the δ parameter tends to the lower bound of the given interval in the optimising process and as such is not equal to zero, but 1×10^{-5}. This difference between the models plays a small role in shifting the confidence interval of the predicted parameters, even though the chosen parameter values are the same. This small role results in slightly narrower intervals than those associated with the Hawkes process model parameters. The optimised model parameters are given in the marked conditional intensity function given below.

$$\lambda(t) = 0.0039 + \sum_{t_i<t} 0.323 \times e^{(1\times10^{-5}\times m_i - 0.409(t-t_i))}$$

A subsection of the plotted marked conditional intensity function for the first dry season is given below. The subsection plot is identical to that of the Hawkes given in figure 14 due to their λ, α and β parameters being equal and the marks impact parameter being approximately zero. Similarly, the marked conditional intensity over the whole interval for the first dry season given by Figure 34 in the Appendix looks identical to the Hawkes given by Figure 33 in the Appendix.

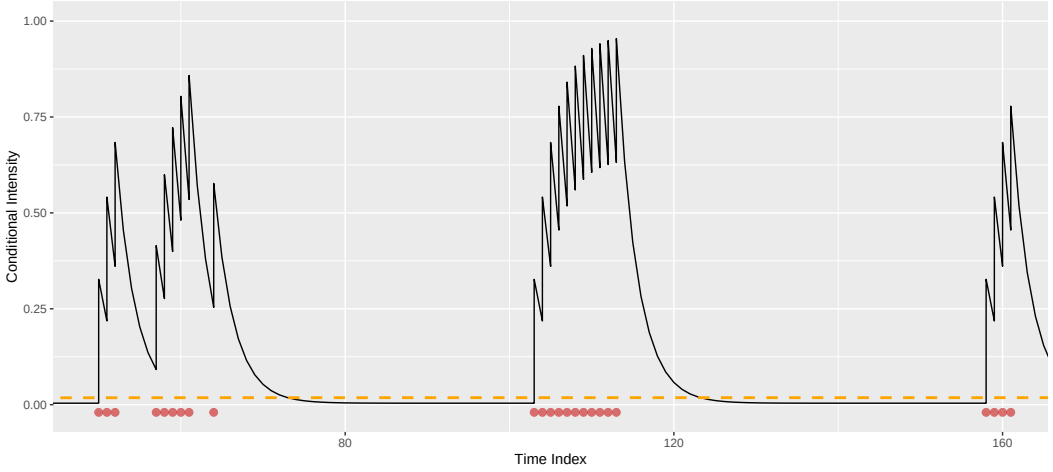

Figure 15: Subsection of dry 1 extreme marked Hawkes conditional intensity with parameters $\lambda = 0.0039$, $\alpha = 0.323$, $\beta = 0.409$ and $\delta = 0$ given by the black line. Events are given by red dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

6.4 Model checking and comparison

In order to compare the model fit for dry extreme 1, the AIC and BIC will be used from equation 24. The negative log-likelihood of the Hawkes model is smaller than the marked Hawkes model (Table 10). This results in both the AIC and BIC being lower for the Hawkes model than for the marked Hawkes model. As the addition of the marked δ parameter does not change the λ, α and β parameters from the Hawkes model to the marked Hawkes model, the added complication of the CO_2 marks may not beneficial. The fit of both the Hawkes and marked Hawkes models is explored further as the best model may still fail to capture the important features present in the data.

Table 10: Maximised log-likelihood for the Hawkes and marked Hawkes models for dry extreme 1 and resulting AIC and BIC measures for model comparison.

Model	**− Log-likelihood**	**AIC**	**BIC**
Hawkes	486.622	979.244	1465.244
Marked Hawkes	1013.961	2035.922	2683.922

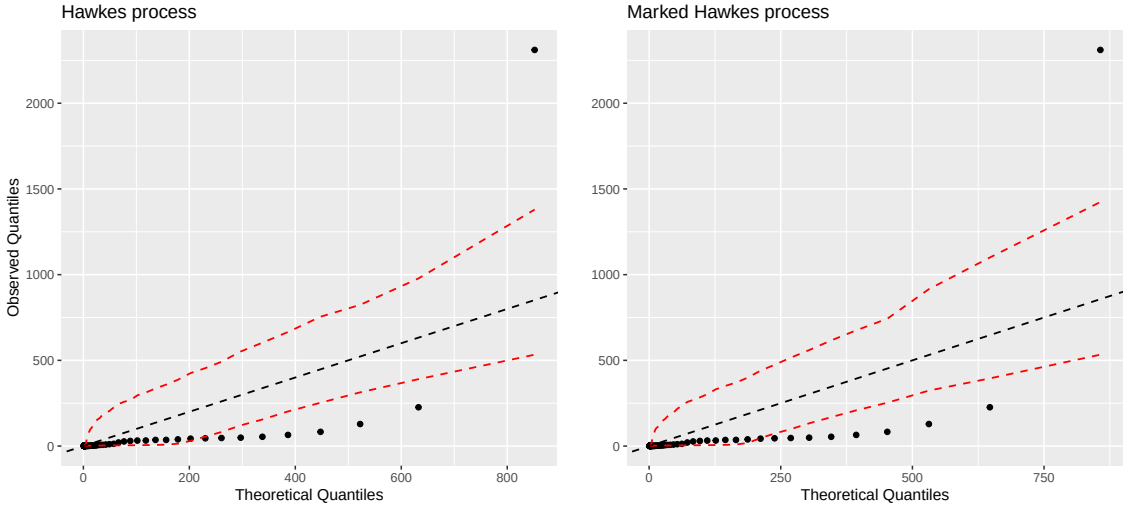

Figure 16: Q-Q plots for the Hawkes and marked Hawkes models for the first dry extremes, plotting the observed inter-arrival times against the reference distribution of inter-arrival times generated by Algorithm 1.

Neither the Hawkes nor the marked Hawkes models fit the CO_2 extreme data well (Figure 16). For both models, most plotted points are well below the 95% simulated confidence interval, indicating that the observed inter-arrival times are lower than those simulated from the proposed models. The Hawkes and marked Hawkes models are underestimating the inter-arrival times between events as most of the points are below the Q-Q line.

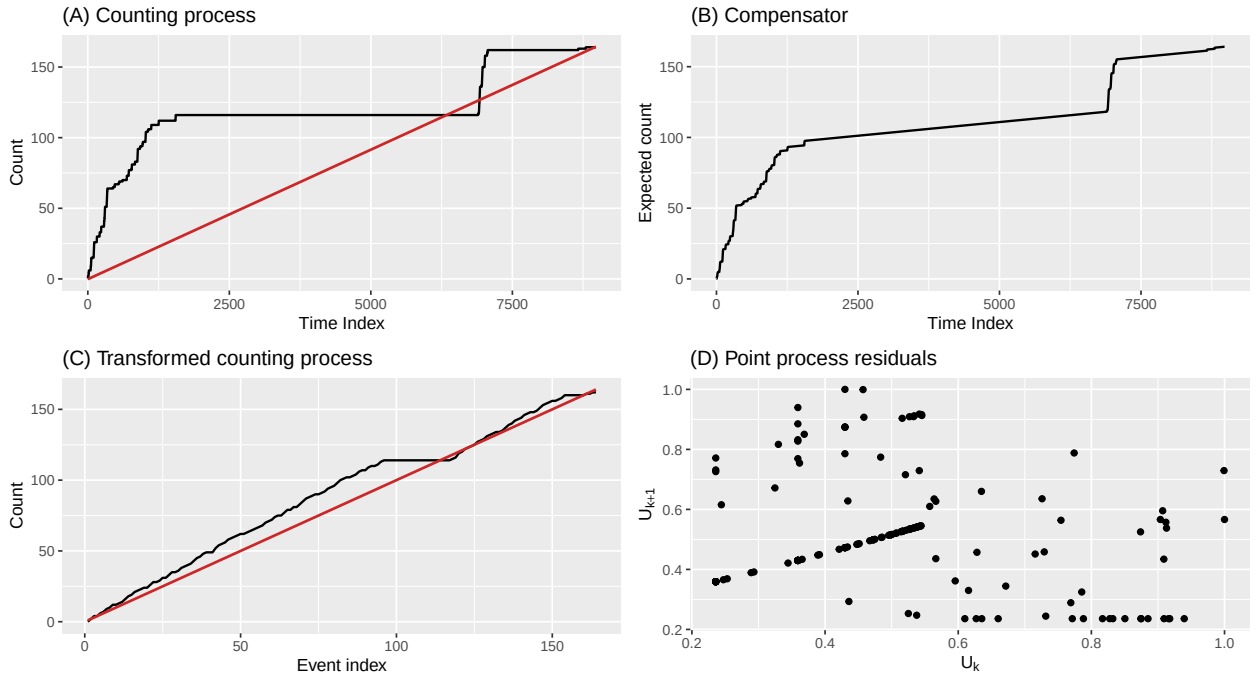

Figure 17: Fit checking plots for the Hawkes process model. (A) The counting process (black line) with the expectation of the counting process (red line), denoted $\mathbb{E}[N(t)]$. (B) The compensator of the Hawkes process. (C) The counting process resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the counting process (red line), denoted $\mathbb{E}[N^*(t)]$. (D) A Qualitative auto-correlation test where $U_k = F(t_k^* - t_{k-1}^*) = 1 - e^{-(t_k^* - t_{k-1}^*)}$.

The above figure gives various model fitting plots for the Hawkes model for the first dry season. Figure 17 (A) plots the observed counting process against the expectation of the counting process, $\mathbb{E}[N(t)]$, given by the first moment of the counting process. This plot demonstrates the intense clustering in the data set. The observed process increases quickly initially before plateauing for a long time interval before the next arrival. The mean of the counting process shows the general trajectory of event occurrence. Both the expectation of the counting process and the observed process give the same number of events at the end of the interval. The plot in Figure 17 (B) gives the expected number of events over the interval and further demonstrates the strong clustering in the CO_2 data. It is clear to see two periods of major increase and an otherwise gradual expected increase in occurrences of events. This shows that, under the Hawkes process model, the expected counting process demonstrates clustering. Figure 17 (C) gives the random-time change transformed counting process against its expectation. It supports the evidence from the plot (A): the Hawkes model does not fit the expected distribution of the inter-arrival times. Figure 17 (D) shows the plot of the U_k values for neighbouring intervals and assessed these values for auto-correlation. Under the model assumptions, the U_k values, which are representative of the inter-arrival times of the

transformed process, are supposed to be i.i.d. uniform random variables. There is some evidence for auto-correlation in the observed data, and as such the transformed inter-arrival times may not be independent and the proposed model does not appear to be a good fit for the data.

The interpretation for figure 18 below is equivalent to the above interpretations, because the Hawkes and marked Hawkes models are equivalent due to their λ, α and β parameters being equal and the marked parameter, δ being approximately equal to zero.

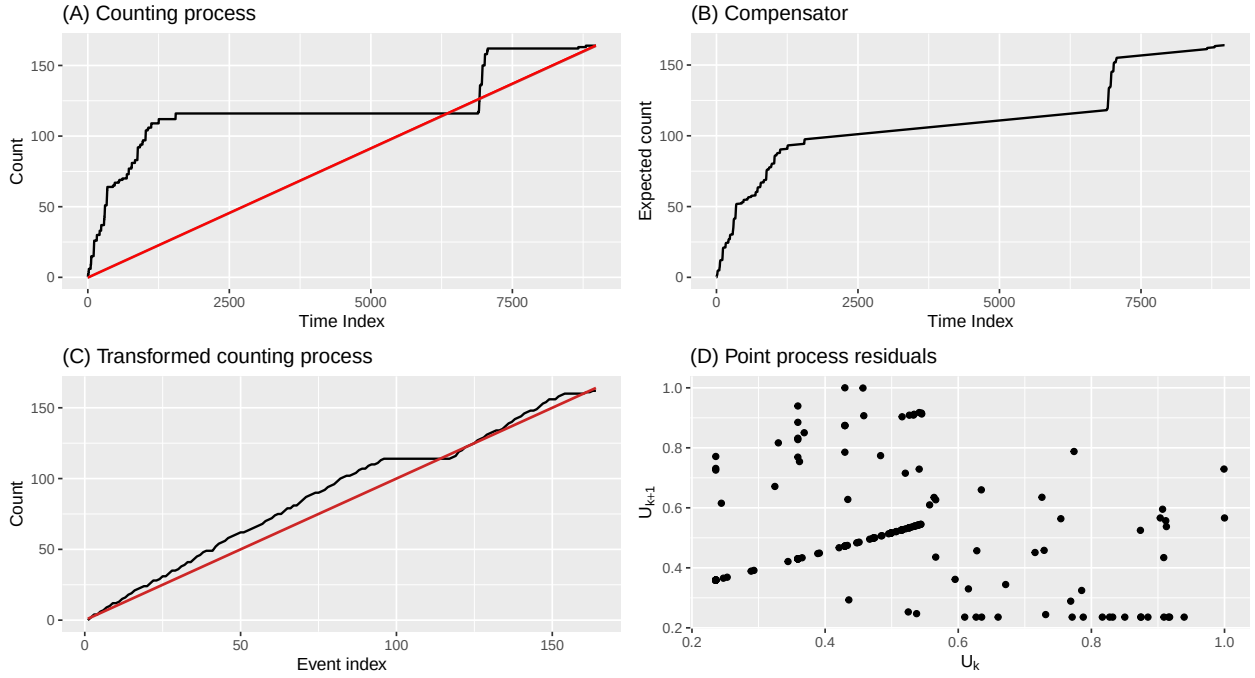

Figure 18: Fit checking plots for the marked Hawkes process model. (A) The counting process (black line) with the expectation of the counting process (red line), denoted $\mathbb{E}[N(t)]$. (B) The compensator of the Hawkes process. (C) The counting process resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the counting process (red line), denoted $\mathbb{E}[N^*(t)]$. (D) A Qualitative auto-correlation test where $U_k = F(t_k^* - t_{k-1}^*) = 1 - e^{-(t_k^* - t_{k-1}^*)}$.

Since the model formulations are so similar between the Hawkes and marked Hawkes process models for the CO_2 extremes of the first dry season, the marked Hawkes process model appears to offer no improvement on the fit to the data and the added complication of including the magnitude of the CO_2 exceedance in the model is not justified. This is demonstrated in Figures 35 and 36 in the Appendix, which show that the difference between the models is negligible.

7 Wet extreme application and results

7.1 Simple Poisson process

The intensity for the simple Poisson process model for the wet extreme is constant at a value of 0.0186. This intensity is invariant over time and independent of the history of the process and can be interpreted as the expected number of events at any given time in the interval being 0.0186. The model formulation for the intensity of the simple Poisson process is given by:

$$\lambda = 0.0186$$

The suitability of the simple Poisson process model must be explored for the wet extremes.

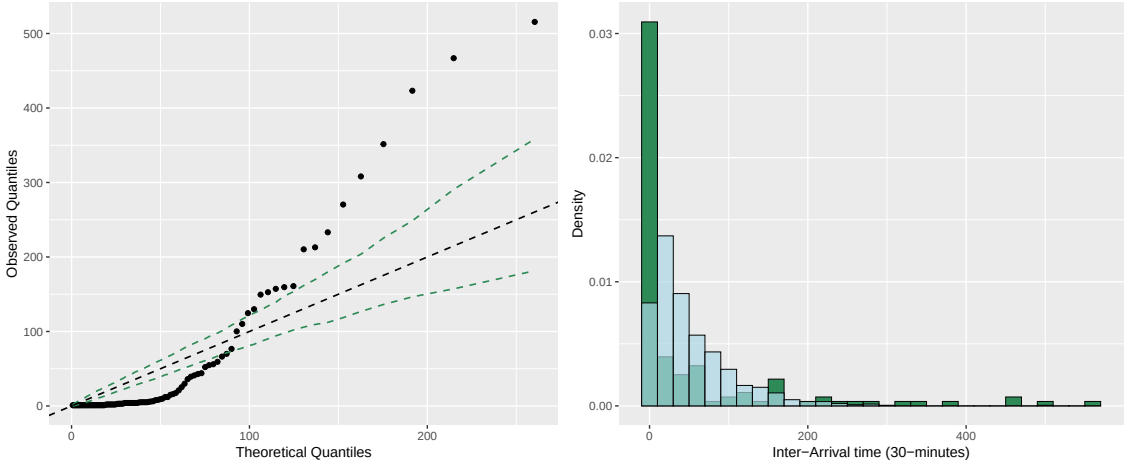

Figure 19: Quantile-quantile plot (left) of the observed inter-arrival times against a theoretical exponential distribution with the rate λ. The black dotted line is the Q-Q line, while the green dotted line gives a 95% simulated confidence interval. The right-hand plot gives the observed distribution of the inter-arrival times (green) against a simulated inter-arrival time sample of size 1000 from an exponential distribution with rate $\frac{1}{\lambda}$ (blue).

As demonstrated by Figure 19, the observed distribution of the inter-arrival times between the CO_2 events does not appear to fit the reference exponential distribution with rate $\frac{1}{\lambda}$. Most of the plotted points on the Q-Q plot fall below the 95% simulated confidence interval, indicating that the Poisson process model does not capture the clustering in the data well, similarly to the first dry extreme. This is supported by the histograms of the inter-arrival time distributions in the right-hand plot (Figure 19). The observed distribution shows a much higher density of very small inter-arrival times than the simulated distribution, showing that the CO_2 events are more closely followed by one another than the proposed model specifies. This is demonstrated by Figure 20, which plots the observed event history against a simulated event history from the reference exponential distribution.

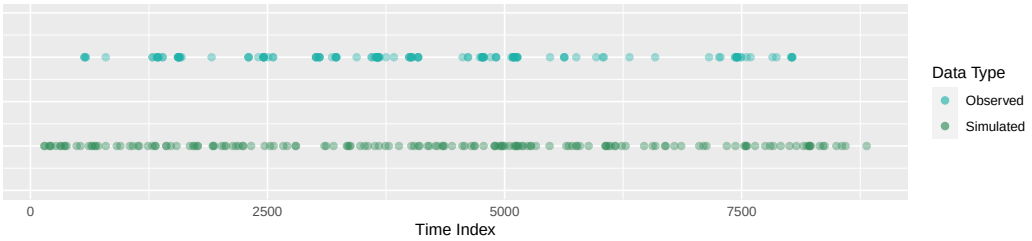

Figure 20: Wet extreme observed events (blue) with simulated events (green) over the same interval. Simulated events via an exponential distribution with rate $\frac{1}{\lambda}$. There are 140 observed events and 164 simulated events.

As seen in the above plot, the simulation overestimated the number of events on the interval of interest by 24. While this is a relatively large overestimation of approximately 14.5%, it is important to note that this is only one sample taken for demonstrative purposes. A true indication of whether the proposed model has a bias for over-estimation of event numbers could be gained through the analysis of many samples. The wet extreme consists of many small clusters with a few CO_2 events in each. It is not as tightly clustered as the dry extreme 1, but there are still visible areas of clustering. The simulated history shows very consistent and regularly-spaced events, with very little evidence of clustering. This plot provides a visual aid in demonstrating the failure of the simple Poisson process to suitably model data which features clusters.

7.2 Hawkes process

As demonstrated by the assessment of the fit of the Poisson process to the data from the wet extreme, a model which can account for clustering in the data may be more suitable to model this system. Given in the table below are the optimised parameters for the Hawkes process model fit to data from the wet extreme. These parameters were found by maximising the log-likelihood found in equation 14. The maximized log-likelihood value is equal to -586.659.

Table 11: Optimised Hawkes parameters for wet extreme with corresponding 95% confidence interval.

Parameter	**Value**	**95% confidence interval**
λ	0.0065	(0.00649, 0.00654)
α	0.141	(0.1414, 0.1416)
β	0.235	(0.2349, 0.2352)

In the wet season extremes, the occurrence of a CO_2 arrival increases the conditional intensity function by 0.141. The rate of decay after the occurrence of an event is 0.235, towards a baseline intensity of 0.0065. The ranges of the associated 95% confidence intervals for the model parameters are 0.00005, 0.0002 and 0.0003 for the λ, α and β parameters, respectively. The Hawkes intensity formula with the associated parameter values is:

$$\lambda(t) = 0.0065 + \sum_{t_i<t} 0.141 \times e^{-0.235(t-t_i)}$$

This Hawkes conditional intensity for wet extreme is plotted over a subsection of the interval which illustrates the excitation effect an event has on the intensity and the decay thereafter (Figure 21). The conditional intensity over the whole wet season interval is given by Figure 37 in the appendix. Both of these figures also demonstrate the less tight-clustered nature of the CO_2 data in the wet season extremes when compared to the dry season extremes.

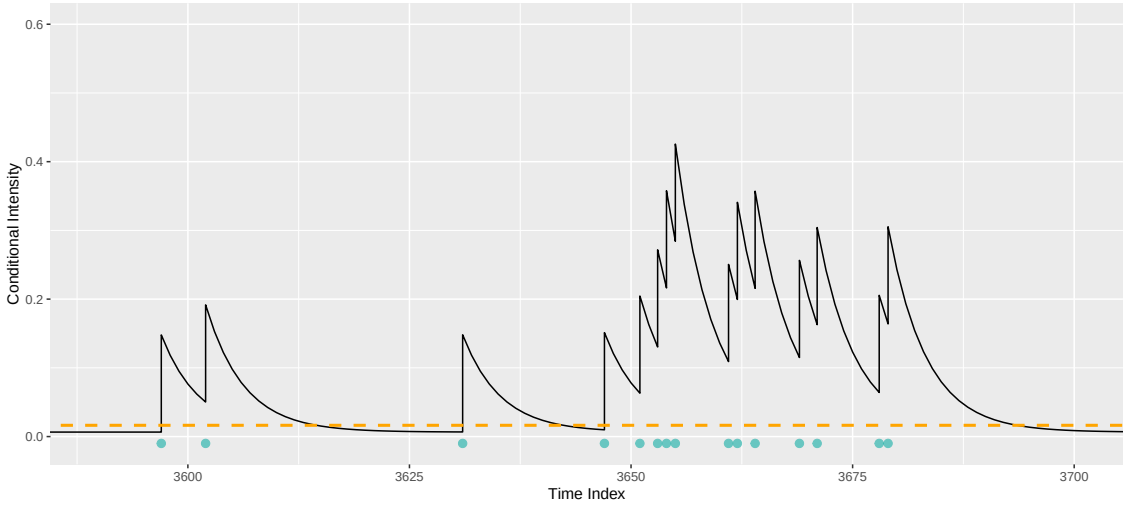

Figure 21: Subsection of wet extreme Hawkes conditional intensity with parameters $\lambda = 0.0065$, $\alpha = 0.141$ and $\beta = 0.235$ given by the black line. Events are given by blue dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

7.3 Marked Hawkes process

The parameters in Table 12 are the optimised parameters for the marked Hawkes process model fit to data from the wet season extremes. These parameters were found by maximising the log-likelihood found in equation 19, which was found to be -1185.693.

Table 12: Optimised Marked Hawkes parameters for dry extreme 1 with corresponding 95% confidence interval.

Parameter	Value	95% confidence interval
λ	0.0065	(0.00647, 0.00653)
α	0.105	(0.099, 0.110)
β	0.230	(0.223, 0.237)
δ	0.0098	(0.0092, 0.0105)

With the occurrence of an event with magnitude mark d, the instantaneous increase of intensity in the system is $0.105 \times (e^{0.0098})^d$, thereafter decaying exponentially at a rate of 0.230 towards the baseline intensity of 0.0065.

The ranges of the 95% confidence intervals for the marked Hawkes parameters for the wet extreme are 0.00006, 0.011, 0.014 and 0.0013 for the λ, α, β and δ parameters respectively. These 95% confidence intervals are wider than for the Hawkes parameters (Table 11), indicating that a wider range of optimal parameters were found.

Some comparisons can be drawn between the Hawkes and marked Hawkes conditional intensity functions for the wet extreme. While the baseline intensity remains the same for both models, the α parameter decreases from 0.141 to 0.105 with the inclusion of the marked parameter, δ, indicating that there is an element of the excitation effect that an event has on the system which is due to the CO_2 magnitude of the event. The β parameter value decreases by 0.005 across the model formulations, indicating that under the marked formulation, the rate of decay of conditional intensity following the event is slightly slower.

$$\lambda(t) = 0.0065 + \sum_{t_i<t} 0.105 \times e^{(0.0098 m_i - 0.230(t-t_i))}$$

Figure 22 demonstrates the marked Hawkes conditional intensity function over a subsection of the interval of interest for the wet season extremes. While Figure 38 in the Appendix is the marked Hawkes conditional intensity over the entire wet season interval.

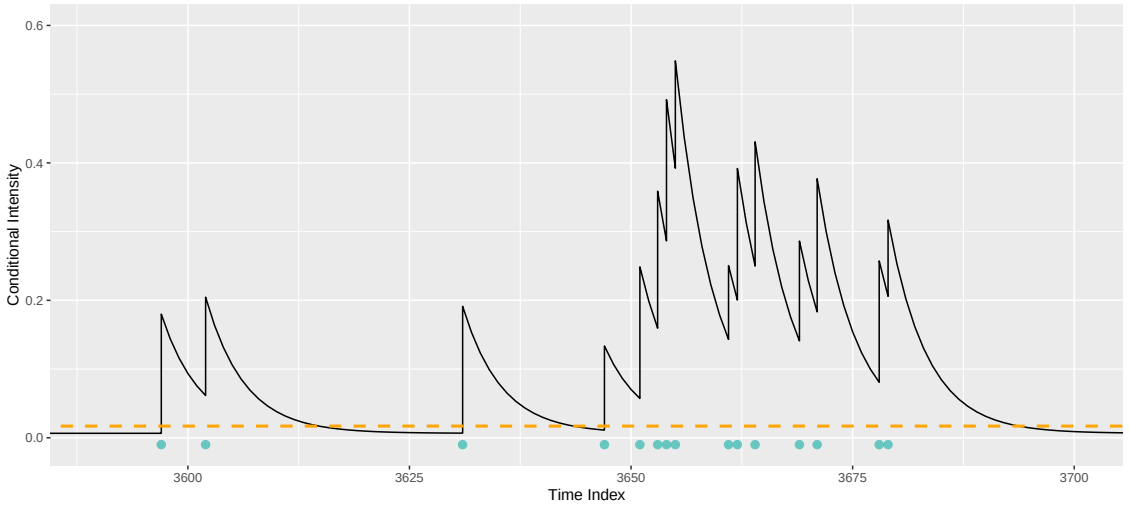

Figure 22: Subsection of wet extreme Marked Hawkes conditional intensity with parameters $\lambda = 0.0065$, $\alpha = 0.105$, $\beta = 0.230$ and $\delta = 0.0098$ given by the black line. Events are given by blue dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

7.4 Model checking and comparison

The negative log-likelihood value is larger for the marked Hawkes model than the Hawkes model which can be attributed to the additional mark parameter δ. According to both AIC and BIC from equation (24), the Hawkes model is more appropriate due to the associated lower AIC and BIC values (Table 13). This indicates that the added complication of the marks may not be necessary. To determine which model fits the data better overall, further checking is performed.

Table 13: Maximised log-likelihood for the Hawkes and marked Hawkes models for wet extreme and resulting AIC and BIC measures for model comparison.

Model	**− Log-likelihood**	**AIC**	**BIC**
Hawkes	586.659	1179.318	1593.318
Marked Hawkes	1185.693	2379.386	2931.386

As demonstrated by the left-hand plot of Figure 23, the Hawkes process model is a good fit for the data, with all of the points falling on or within the 95% simulated confidence interval. This indicates that the observed data could reasonably be expected to have come from the reference distribution under the Hawkes model. The right-hand plot of Figure 23 gives the Q-Q plot testing fit for the marked Hawkes model and shows almost no improvement in fit in comparison to that of the Hawkes process model. This indicates that the addition of the marked parameter associated with the magnitude of the CO_2 events has almost no effect on model fit and may not be necessary.

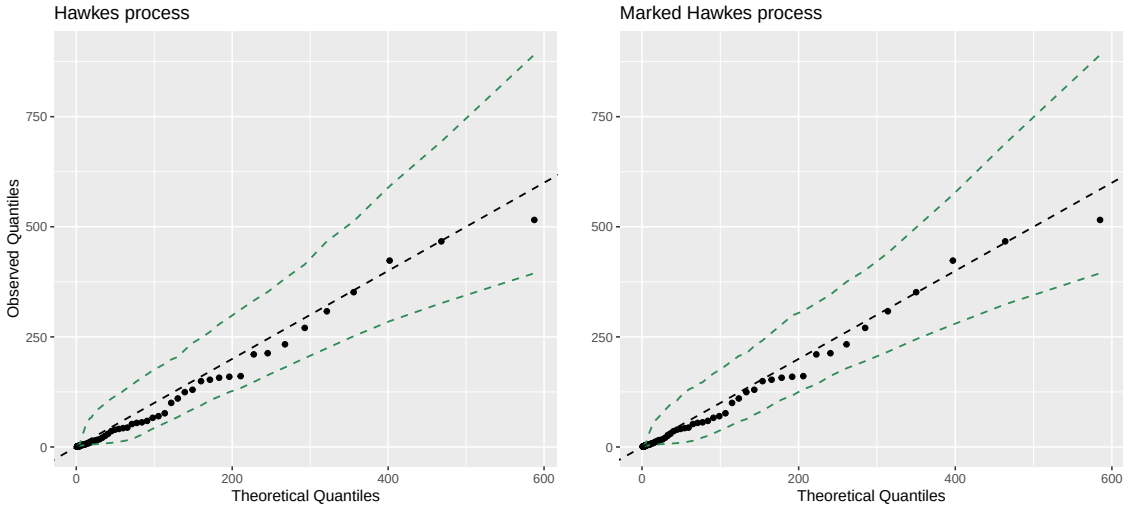

Figure 23: Q-Q plots for the Hawkes and marked Hawkes model for the wet extremes, plotting the observed inter-arrival times against the reference distribution of inter-arrival times generated by Algorithm 1.

The plots in Figure 23 suggest that both the Hawkes and marked Hawkes process models are well fitted to the data, with the marked Hawkes process model appearing to have an very slightly better fit to the CO_2 wet data than the Hawkes process model. Further model fit checks need to be performed on the Hawkes and marked Hawkes process models for the wet season extremes.

Figure 24 (A) shows that the observed counting process for the wet extremes demonstrates small areas of extreme CO_2 event clustering across the interval. The mean of the counting process and the observed counting process predict the same amount of total extreme CO_2 events across the interval. The plot in Figure 24 (B) gives the expected count of events over the interval. This plot supports the findings of Figure 24 (A), showing small areas of temporal event clustering. Figure 24 (C) gives the random-time change transformed counting process against its expectation. Under the model assumptions, this transformed counting process is expected to follow that of a unit rate Poisson process, which is given by the blue line. The transformed counting process is well-fitted to the expected line, showing a slight deviation around the 100_{th} event index. The plot in Figure 24 (D) indicates that the assumption of independence in the transformed U_k variables is well-supported, with the given plot showing little evidence of auto-correlation between U_k variables.

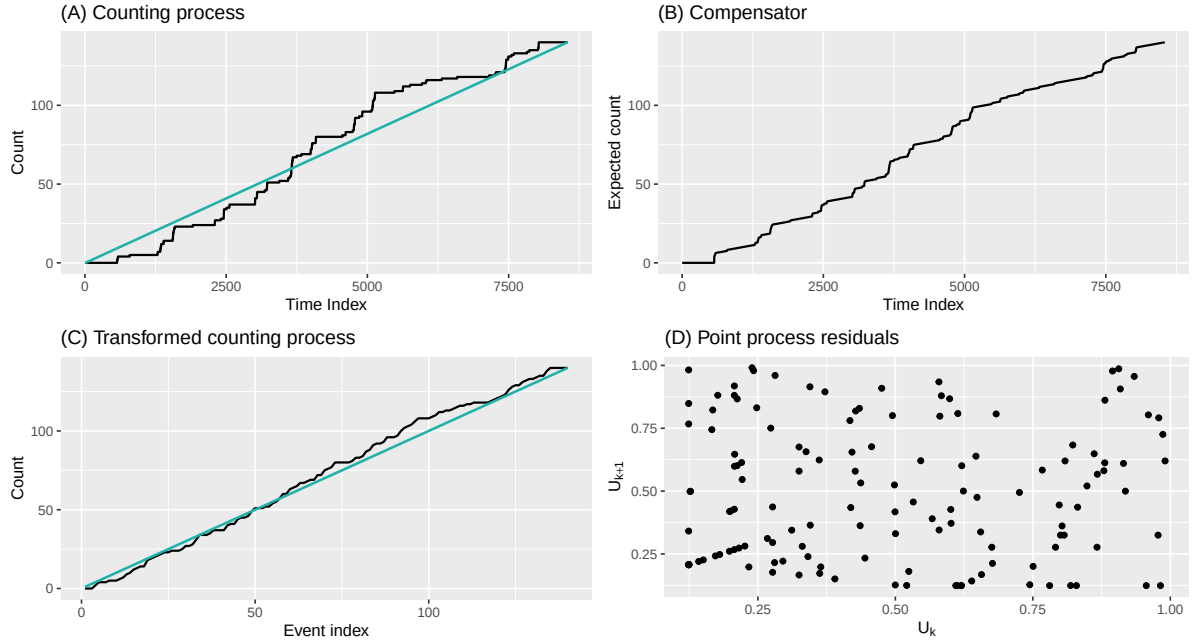

Figure 24: Fit checking plots for the Hawkes process model. (A) The counting process (black line) with the expectation of the counting process (blue line), denoted $\mathbb{E}[N(t)]$. (B) The compensator of the Hawkes process. (C) The counting process resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the counting process (blue line), denoted $\mathbb{E}[N^*(t)]$. (D) A Qualitative auto-correlation test where $U_k = F(t_k^* - t_{k-1}^*) = 1 - e^{-(t_k^* - t_{k-1}^*)}$.

Figure 25 below gives the relevant model-fit check plots for the marked Hawkes model and can be interpreted in the same manner as above. Figure 25 (A) shows the counting process following a similar trajectory to the expectation of the counting process under the marked Hawkes model. It is worth noting, however, that the expectation of the counting process does appear to overestimate the number of events expected in the interval slightly. The compensator plot of the marked Hawkes process given in Figure 25 (B) indicates the small areas of event clustering that are expected across the interval. Similarly to the Hawkes process model, the transformed counting process of the marked Hawkes process given in Figure 25 (C) is a good fit for the expected transformed counting process, indicating that the marked Hawkes process model is a good fit for the wet CO_2 data. The auto-correlation plot of U_k against U_{k+1} given in Figure 25 (D) demonstrates independence between the U_k variables, supporting the fit of the marked Hawkes process model to the wet CO_2 data.

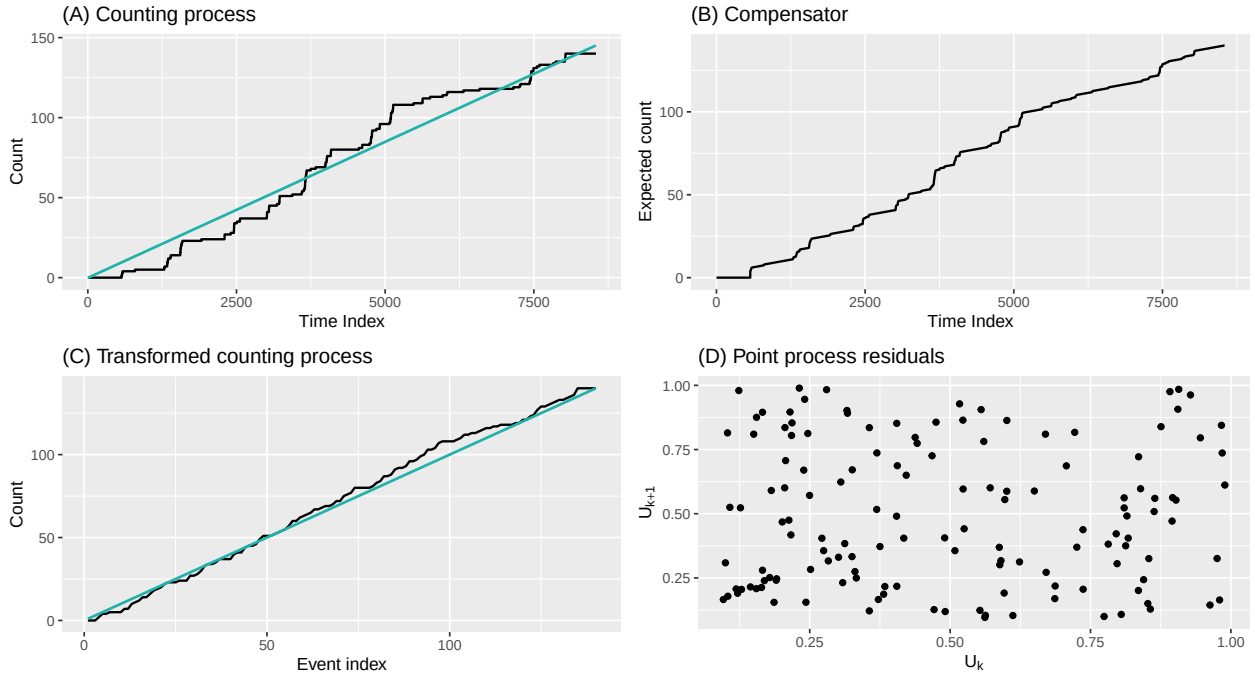

Figure 25: Fit checking plots for the marked Hawkes process model. (A) The counting process (black line) with the expectation of the counting process (blue line), denoted $\mathbb{E}[N(t)]$. (B) The compensator of the Hawkes process. (C) The counting process resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the counting process (blue line), denoted $\mathbb{E}[N^*(t)]$. (D) A Qualitative auto-correlation test where $U_k = F(t_k^* - t_{k-1}^*) = 1 - e^{-(t_k^* - t_{k-1}^*)}$.

Figures 39 and 40 in the Appendix contrast the sub-section of the conditional intensities for the Hawkes and Marked Hawkes process models for the extremes of the wet season, as well as the plot of the counting process and its corresponding mean, and the transformed

counting process and its expectation. These are useful when comparing the model fit to the wet season extremes. Figure 39 shows the variable excitation affects the magnitude of the CO_2 exceedance in the marked Hawkes process model. Looking at the left-hand plot in Figure 40, the expectation of the Hawkes counting process estimates the same number of extreme CO_2 events as were observed in the interval, while the expectation of the marked counting process over-estimates the number of events in the interval. Looking at the plot on the right in Figure 40, the transformed counting process of the marked Hawkes model appears to be slightly better fitted to the unit-rate process than the Hawkes model, though the difference is extremely slight. This is supported by the findings of the Q-Q plots in Figure 23, which show a very minor improved fit for the marked Hawkes model and plot (D) of figures 24 and 25, with the auto-correlation scatter-plot for the marked Hawkes process model being a closer representation of a random scatter.

8 Dry extreme 2 application and results

8.1 Simple Poisson process

The intensity of the simple Poisson process is constant at a value of 0.0104 for the second dry season extremes. This intensity is invariant over time and independent of the history of the process and can be interpreted as the expected number of events at any given time in the interval. The formulation of the intensity for the simple Poisson process model of the second dry season extremes is given by:

$$\lambda = 0.0104$$

The fit and suitability of the simple Poisson process need to be accessed, which can be done through simulation. Figure 26 give plots evaluating the fit of the simple Poisson process to the second dry season extremes to assess if the simple Poisson process is a suitable model.

The majority of the plotted points in the Q-Q plot fall below the 95% simulated confidence interval (Figure 26). In a similar manner to the first dry season extremes, the simple Poisson process tends to over-estimate the inter-arrival times between extreme events of the second dry season and in doing so, fails to capture the clustering of extreme events observed in the second dry season. This finding is supported by the right-hand plot of Figure 26, which gives the density of the observed inter-arrival times against those simulated from a sample of inter-arrival times from the reference distribution. The bulk of the distribution of observed inter-arrival times is very small, while the distribution of theoretical inter-arrival times under the model assumptions shows a much higher density of larger inter-arrival times, providing further evidence that the string clustering in the data is not captured by the Poisson model.

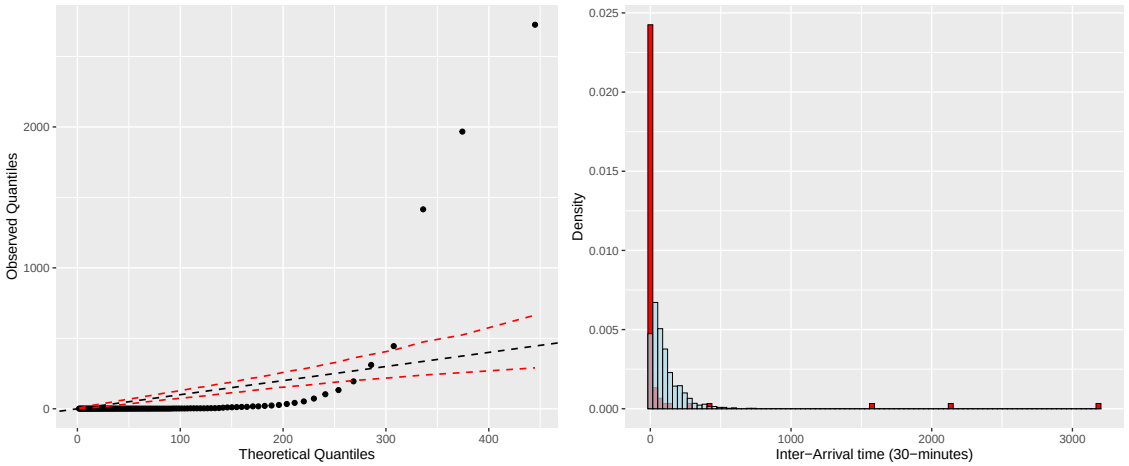

Figure 26: Quantile-quantile plot (left) of the observed inter-arrival times against a theoretical exponential distribution with the rate λ. The black dotted line is the Q-Q line, while the red dotted line gives a 95% simulated confidence interval. The right-hand plot gives the observed distribution of the inter-arrival times (red) against a simulated inter-arrival time sample of size 1000 from an exponential distribution with rate $\frac{1}{\lambda}$ (blue).

To illustrate this finding, the observed history of events is plotted against a simulated history, generated from the assumed distribution of arrival times under the Poisson model.

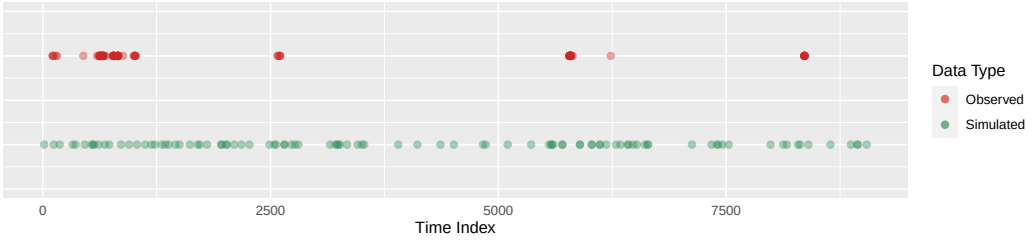

Figure 27: Dry 2 extreme observed events (red) with simulated events (green) over the same interval. Simulated events via an exponential distribution with rate λ. There are 87 observed events and 98 simulated events.

The simulated history overestimated the events by approximately 12.5%, but this is the result of one sample and would be expected to vary between samples. As with the first dry extreme, the observed history is heavily clustered over the interval, while the simulated history is much more regularly spaced, with less clear clustering of events. This demonstrates the need for a model which can account for the clustering and dependence between events when modelling the extremes of the second dry season. The Hawkes process model can account for event clusters by taking the history of events into account when calculating the conditional intensity of the process.

8.2 Hawkes process

Given in Table 14 are the optimised parameters for the Hawkes process model fit to the extremes of the second dry season. These values were found by maximising the log-likelihood found in equation (14), which was found to be equal to -274.6863.

Table 14: Optimised Hawkes parameters for dry extreme 2 with corresponding 95% confidence interval.

Parameter	Value	95% confidence interval
λ	0.0015	(0.0012, 0.0019)
α	0.184	(-0.006, 0.373)
β	0.218	(0.001, 0.435)

Under the Hawkes process model formulation for the extremes of the second dry season, the occurrence of a CO_2 event increases the conditional intensity instantaneously by 0.184. Thereafter, the conditional intensity decays exponentially at a rate of 0.218 towards the baseline intensity of 0.0015. The formulation of the Hawkes process with the optimised parameters is given by:

$$\lambda(t) = 0.0015 + \sum_{t_i<t} 0.184 \times e^{-0.218(t-t_i)}$$

The ranges of the 95% confidence intervals of the parameters are 0.0007, 0.379 and 0.434 for λ, α and β respectively. As α has to be greater than zero, the lower bound for the parameter can be considered approximately equal to 0.

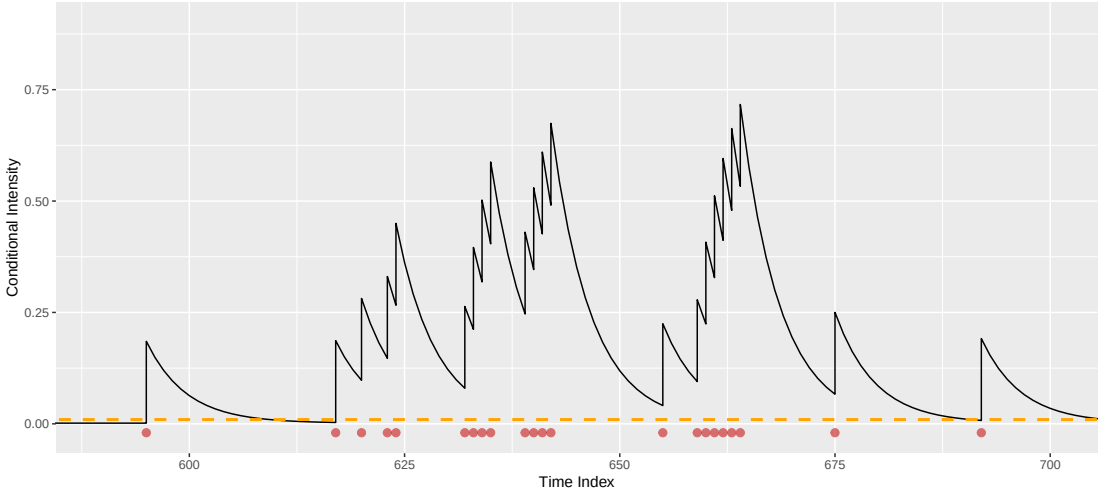

Figure 28: Subsection of dry 2 extreme Hawkes conditional intensity with parameters $\lambda = 0.0015$, $\alpha = 0.184$ and $\beta = 0.218$ given by the black line. Events are the red dots and the orange dotted line is the intensity expectation, $\mathbb{E}[\lambda(t)]$.

The excitation effect of an event on the system is clearly observed, as well as the decay towards the baseline intensity (Figure 28). The clustering of the event occurrences is clearly visible in this subsection plot. The Hawkes conditional intensity plotted over the entire interval for the second dry season can be found in Figure 41 in the Appendix.

8.3 Marked Hawkes process

Table 15 contains the optimised parameters for the marked Hawkes process model fit to extremes from the second dry season. These parameters were found by maximising the log-likelihood found in equation 19, which was found to be -635.145.

Table 15: Optimised Marked Hawkes parameters for dry extreme 2 with corresponding 95% confidence interval.

Parameter	**Value**	**95% confidence interval**
λ	0.0016	(0.0013, 0.0018)
α	0.169	(0.057, 0.281)
β	0.223	(0.119, 0.327)
δ	0.0046	(-0.004, 0.013)

The marked conditional intensity for the CO_2 extremes of the second dry season increases instantaneously by $0.169 \times e^{0.0046(1)}$ with the occurrence of the event with a magnitude mark of 1, before decaying at a rate of 0.223 towards the baseline intensity of 0.0016. The occurrence of some event with a magnitude mark equal to d would result in an instantaneous increase in the intensity of $0.169 \times (e^{0.0046})^d$, thereafter decaying exponentially at a rate of 0.223 to the baseline intensity of 0.0016. The ranges of the 95% confidence interval for the optimised parameters are 0.0005, 0.224, 0.208 and 0.017 for the λ, α, β and δ parameters respectively. The formulation of the marked Hawkes intensity with the optimised parameters is given by:

$$\lambda(t) = 0.0016 + \sum_{t_i<t} 0.169 \times e^{(0.0046 m_i - 0.223(t-t_i))}$$

The baseline intensity increased by 0.0001 with the inclusion of the marked parameter δ. The α parameter value decreases from 0.184 in the Hawkes process to 0.169 in the marked Hawkes process, indicating that some of the excitation effects that an event has on the system can be attributed to the magnitude of the event. The decay term increases by 0.015 from 0.218 in the Hawkes model to 0.223 in the marked Hawkes model, though the rate of decay in the marked model is variable and does not relate to the magnitude of the marks. The subsection of the marked Hawkes intensity shows the excitation effect an event has on the system and demonstrates the clustering of the events over the time interval (Figure 29).

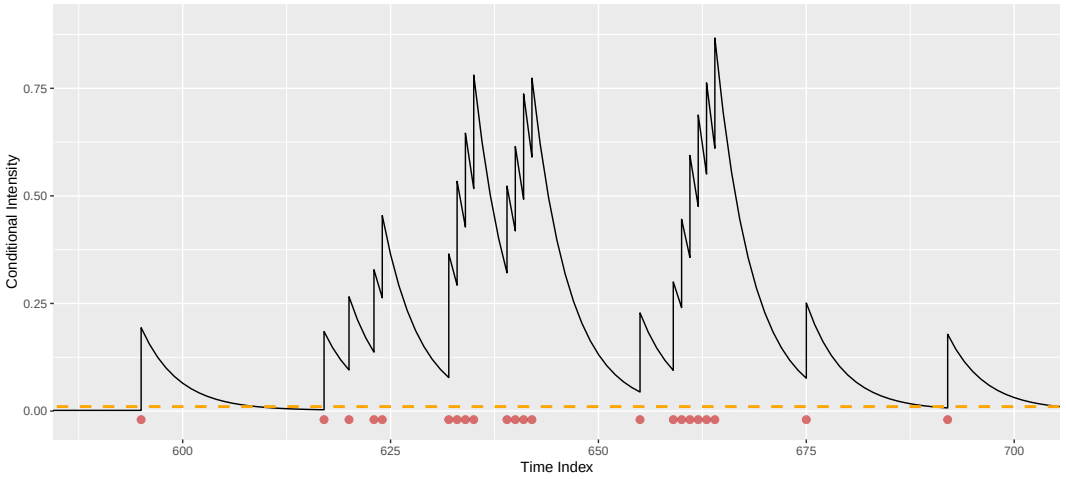

Figure 29: Subsection of dry 2 extreme Marked Hawkes conditional intensity with parameters $\lambda = 0.0016$, $\alpha = 0.169$, $\beta = 0.223$ and $\delta = 0.0046$ given by the black line. Events are given by red dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

8.4 Model checking and comparison

The negative log-likelihood value is larger for the marked Hawkes model than the Hawkes model which can be attributed to the additional mark parameter δ. In terms of both AIC and BIC from equation (24), the Hawkes model may be a better fit due to the lower associated AIC and BIC values (Table 16). This indicates that the added complication of the marks may not lead to significant improvement in the model. To determine which model fits the dry 2 data better overall, further checking is performed.

Table 16: Maximised log-likelihood for the Hawkes and marked Hawkes models for dry extreme 2 and resulting AIC and BIC measures for model comparison.

Model	**− Log-likelihood**	**AIC**	**BIC**
Hawkes	274.686	555.373	810.373
Marked Hawkes	635.145	1278.289	1618.289

The Q-Q plots for the Hawkes and marked Hawkes process models for the second dry season extremes show that the Hawkes process model fit is relatively good, with all of the plotted quantile points falling on or within the 95% confidence interval (Figure 30). Although, it is worth noting that the plotted points do deviate quite strongly from the reference line, indicating that there is still some tendency to overestimate the inter-arrival times. The marked Hawkes Q-Q plot shows no improvement of fit to the observed CO_2 extremes in the second dry season than that of the Hawkes process model, indicating that the addition of a CO_2 magnitude mark parameter to the model does not make a more representative model of the data.

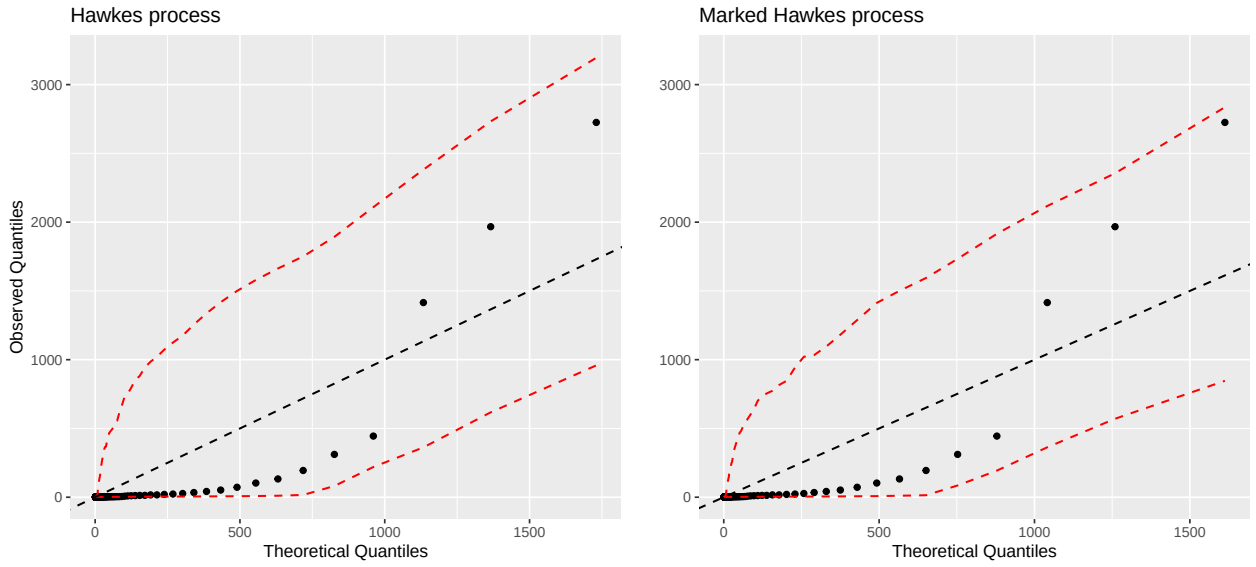

Figure 30: Q-Q plots for the Hawkes and marked Hawkes model for the second dry extremes, plotting the observed inter-arrival times against the reference distribution of inter-arrival times generated by Algorithm 1.

Figure 31 gives further model fit assessment plots for the Hawkes process fitted to the extremes of the second dry season. Figure 31 (A) shows that the observed counting process for the second dry season extremes shows strong clustering of events where the count increases sharply. The analytical mean of the counting process gives the expected trajectory of the counting process. The expected and observed counting processes result in a similar number of events, indicating that the Hawkes process model estimates a similar number of events over the interval as was observed.

The plot in Figure 31 (B) gives the expected count of events over the interval. This plot supports the findings of Figure 31 (A), showing that the expected count of events is strongly clustered under the Hawkes process formulation. Figure 31 (C) gives the random-time change transformed counting process against its expectation. Under the model assumptions, this transformed counting process is expected to follow that of a unit rate Poisson process, which is given by the red line. The transformed counting process is close to, but rarely exactly fits the expected line in this plot, and is indicative that there is room for improvement in terms of the fit of the Hawkes process model to the extremes of the second dry season. This observation is supported by Figure 31 (D), which shows evidence for a degree of auto-correlation between consecutive values in the residual process, indicating that the assumption of the U_k variables being i.i.d. uniformly distributed may not stand under this model formulation.

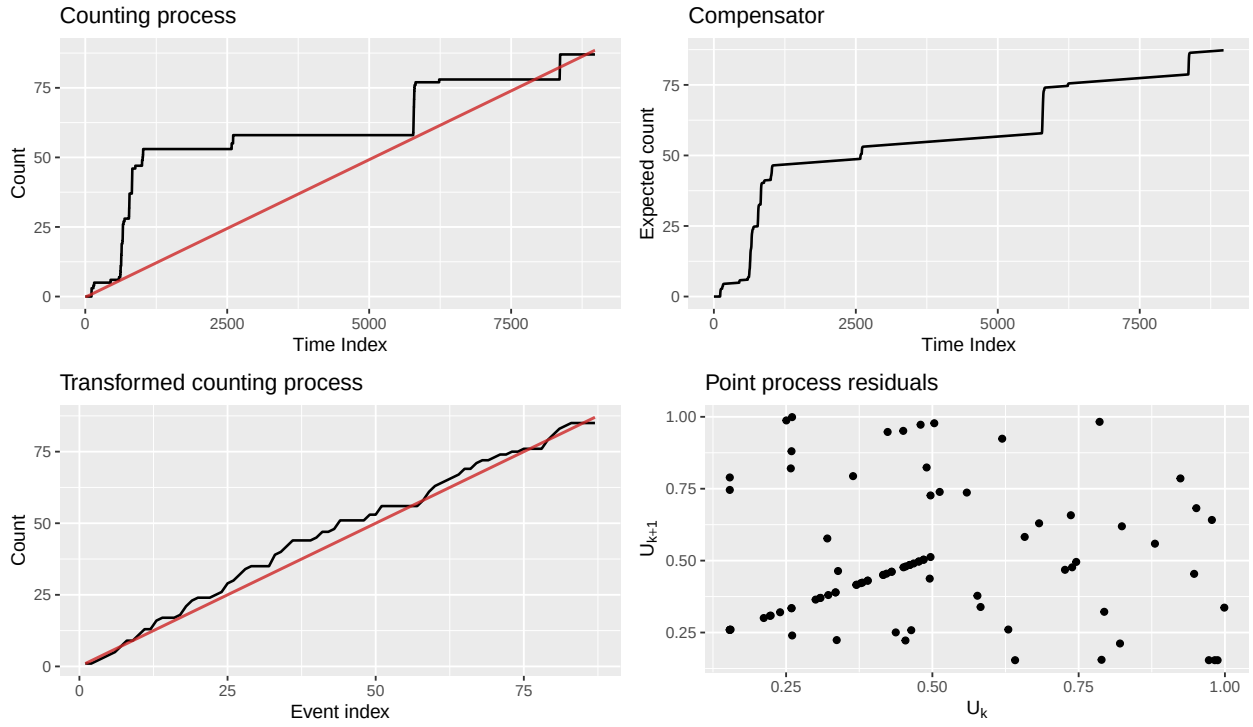

Figure 31: Fit checking plots for the Hawkes process model. (A) The counting process (black line) with the expectation of the counting process (red line), denoted $\mathbb{E}[N(t)]$. (B) The compensator of the Hawkes process. (C) The counting process resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the counting process (red line), denoted $\mathbb{E}[N^*(t)]$. (D) A Qualitative auto-correlation test where $U_k = F(t_k^* - t_{k-1}^*) = 1 - e^{-(t_k^* - t_{k-1}^*)}$.

Figure 32 gives some model fit checking plots for the marked Hawkes process model of the extremes of the second dry season. Plot (A) shows that the expectation of the counting process for the marked model appears to overestimate the number of events in the interval. Plot (B) shows that the expected count of events over the interval does show clustering under this model, with periods of sharp and gradual increases in expected count. The transformed counting process in the plot (C) indicates that the marked Hawkes process model is a fairly good fit for the data, with the transformed counting process line following the expected unit-rate counting process line relatively well, with a small deviation towards the middle of the event index. The auto-correlation plot of the residual process (D) shows that there is still a degree of dependence that can be observed between some points in the residual process.

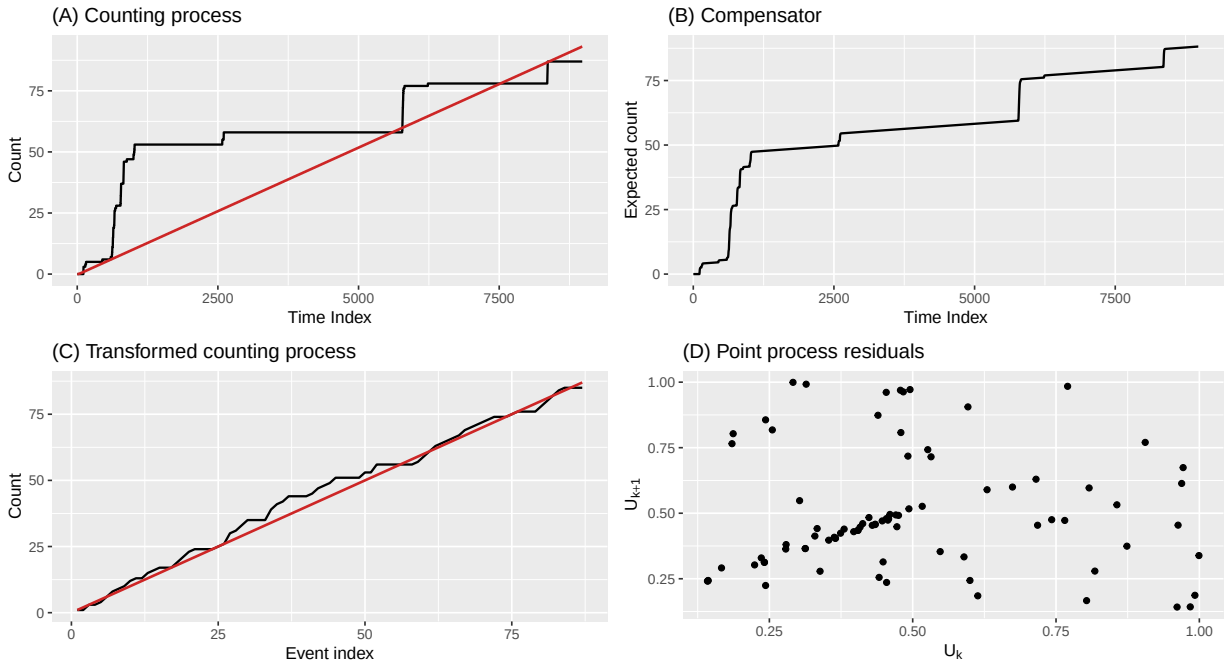

Figure 32: Fit checking plots for the marked Hawkes process model. (A) The counting process (black line) with the expectation of the counting process (red line), denoted $\mathbb{E}[N(t)]$. (B) The compensator of the Hawkes process. (C) The counting process resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the counting process (red line), denoted $\mathbb{E}[N^*(t)]$. (D) A Qualitative auto-correlation test where $U_k = F(t_k^* - t_{k-1}^*) = 1 - e^{-(t_k^* - t_{k-1}^*)}$.

Some comparisons can be drawn between the model fit testing plots for the Hawkes process and the marked Hawkes process models for the extreme CO_2 events in the second dry season. A figure comparing the plots for both models of the mean of the counting process and the transformed counting process against its expectation can be found in Figure 44 in the Appendix. The left-hand plot of the counting process with its mean demonstrates that under the marked Hawkes model, the mean of the counting process is over-estimated, while under the Hawkes model it is slightly under-estimated. In the right-hand plot of Figure 44, the transformed counting process of the marked Hawkes process model is a slightly closer approximation of the expectation of the transformed counting process than that of the Hawkes process model, indicating that the marked Hawkes process model is slightly better in this respect, but the difference is minor and neither model is particularly well-fitted to the data based on this test. This is supported by the Q-Q plots assessing the distribution of the inter-arrival times given in Figure 30, in which any difference in fit between models is very slight. The auto-correlation scatter-plots given in plot (D) of Figures 31 and 32 indicate that the marked Hawkes model is a slightly better fit for the CO_2 extremes of the second dry season, as the plot associated with the marked model is a closer approximation to random scatter, however there is evidence against independence in both plots.

9 Discussion

The simple Poisson process model structure assumes that the inter-arrival times between extreme CO_2 events in each season are independent and exponentially distributed with a rate of the inverse of the mean inter-arrival time in the given season. This assumption was found to not stand for any of the seasons, with the Poisson model showing a strong bias for over-estimation of the inter-arrival times between extreme CO_2 events. This overestimation can be seen in all of the Poisson process Q-Q plots as the Q-Q line is higher than the bulk of the points. The overestimation is not surprising, given that in the exploration of the data, extreme CO_2 events were found to occur in clusters temporally in all of the seasons. Thus, the constant intensity modelled by a Poisson point process is not a suitable model for this clustered data set, as the expected number of extreme CO_2 events is variable over the interval, with an extreme CO_2 event being more likely to occur in the small time period following an event. Given these findings, a model that can represent the clustering of events which have a self-exciting nature may be a more suitable model for this data set and environmental time series in general.

Hawkes process models were fit to the extreme CO_2 readings of each season respectively. The self-exciting Hawkes process model was chosen as, due to the clustering found in the data, the occurrence of an event should have an excitation effect on the conditional intensity given by the model. The Hawkes process model fit in the first dry season is mildly improved on the basic Poisson model fit, but still shows a strong bias for over-estimation of the extreme CO_2 event inter-arrival times. This bias is clearly demonstrated in Figure 16. This tendency for over-estimation of event inter-arrival times in the first dry season is likely due to the combination of the very tight clustering of events observed in the season with the few long time periods between consecutive events. These very long inter-arrival times drive the average inter-arrival time up, making it not truly representative of the system.

The Hawkes process models for the CO_2 extremes in both the wet and second dry seasons show a clear improvement on the basic Poisson model. The Hawkes process model for the wet season extreme CO_2 readings in particular appears to be a good fit for the underlying distribution, as shown by the left-hand plot of Figure 23. In both the wet season and the second dry season, the Hawkes process model is less prone to over-estimation of extreme event inter-arrival times than the Poisson model, indicating that the Hawkes model is more able to take temporal event clustering into account, which is essential when aiming to capture the salient features of this data set. It is worth noting that the intense clustering of extreme CO_2 events observed in the dry seasons appears challenging for this formulation of the Hawkes process to capture; while the tendency for over-estimation of inter-arrival times for these seasons was improved on the Poisson process, the Hawkes process still demonstrated this same bias, although to a lesser degree.

Testing the temporal element of the CO_2 extremes of the wet and second dry seasons using the random time change theorem, indicated that although the Hawkes fit was improved on that of the Poisson process, the fit can still be improved further. This was clearly demonstrated

by the auto-correlation test of the U_k variables, which showed some evidence for dependence.

The addition of a CO_2 exceedance magnitude mark was investigated as a possible mechanism for modelling an underlying state in the system, which may be captured by the magnitude of the exceedance, that may further explain the observed history of events. This led to the formulation of a marked Hawkes process model for the CO_2 extremes of each season. Interestingly, the parameter associated with the magnitude marks, δ, tended to zero in the optimisation process for the CO_2 extremes of the first dry season. It appears that the inclusion of a mark is not a necessary addition to model the extremes of this season and the added complication thereof is not justified. The marked parameter for the CO_2 extremes of the second dry season, however, did not tend to zero and the addition of the CO_2 magnitude marks was found to play a role in the intensity. This discrepancy between the dry seasons can be understood by looking at the relative distribution of the CO_2 exceedance marks for the first and second dry season extremes respectively, given in Figure 10. The exceedance marks of the first dry season in comparison to the second show a mark distribution with lower mark values on average and significantly less spread. The marks of extreme events in the first dry season are fairly homogeneous and its events are very tightly clustered, resulting in the effect of the mark of the events being diminished, as much of the excitation in the system can be captured by the alpha parameter due to the very high concentration of consecutive events.

For all of the seasons modelled, the marked model formulations only showed very slight improvements in model fit, if any. For the first dry season, the addition of CO_2 magnitude mark did not play a role in the conditional intensity, indicating that the addition of a marked parameter into the model was entirely unnecessary. For the wet and second dry season extremes, the addition of the CO_2 magnitude mark hardly indicated any improved fit. These findings are supported by the associated AIC and BIC values for each of the Hawkes and marked Hawkes models, with AIC and BIC values for all of the marked models being greater than those associated with the Hawkes models, indicating that any improved fit gained is not be a good trade-off for the large increase in model complexity that occurs with the inclusion of the marked parameter. It is worth noting that for the wet and second dry season, the marked models were slightly better performing regarding the temporal testing of the data which was done using the random time change theorem. The $N^*(t)$ process for the marked Hawkes models was a slightly closer approximation to the expectation of $N^*(t)$ than those of the Hawkes models and the auto-correlation plots of the U_k variables were more representative of random scatter for the marked models than the Hawkes models of either season. However, these improvements are comparatively insignificant when considering the added complexity of the inclusion of the marked parameters.

Given that the inclusion of marks provides very little, if any, improvement to model fit, the Hawkes process model parameters can be compared between the seasons to gain an understanding of how the underlying system of extreme CO_2 values operate in each season. The baseline intensity, λ of the wet season is much larger than that of either dry season. This indicates that the wet season is more likely to observe an event when the excitation effect

of any previous events has dissipated. This observation is supported by the looser clustering of the wet season extreme events. Because events are not as tightly clustered as in the dry seasons, an event is comparatively more likely to occur without the occurrence of a previous event in recent history. The α parameter is higher for the extreme CO_2 values of the first dry season than either of the other two seasons. This is also supported by the strong clustering of extreme CO_2 events observed in the first dry season, resulting in the instantaneous increase in intensity that an event induces in the system should be greater than in any other season, as events are more likely to follow directly after one another. The parameter related to the rate of conditional intensity decay, β, is also larger in the first dry season than in any other. This can be understood by considering a previously mentioned constraint placed on the system such that $\beta > \alpha$ prevents the 'explosion' of the point process. The extreme CO_2 events are so tightly temporally clustered that the instantaneous increase in intensity is large. The corresponding rate of decay is also large which accounts for the fact that events are more likely to occur consecutively than any later, thus the chance of observing an event drops very sharply as the system time progresses after the event.

Overall point process models appear to be a suitable choice for modelling an environmental time series consisting of CO_2 observations. Point processes are particularly useful for modelling extreme events, due to there being natural clustering between extreme events in the environment with the potential for events to excite another event. The point processes that were used to model the CO_2 extremes were the simple Poisson process, the Hawkes process with an exponential decay function and the marked Hawkes process with exponential decay and impact functions. Most literature concerning point processes does not discuss and compare the three aforementioned models but rather focuses on one, hence this paper compares the models as well as comments on the suitability of a point process model for environmental time series.

A limitation of this analysis was that the log-likelihood surface is very complex, likely with multiple local maxima, so there is a degree of uncertainty surrounding whether or not the optimised parameters for both the Hawkes and marked Hawkes models are at the true global maximum for the log-likelihood function. Although a grid-like search was used to increase the chances of convergence to a local maximum, one cannot be certain that the final selected parameters correspond to the global maximum. As such, an alternative direct numerical maximisation (DNM) routine for finding the parameter should be considered as a check that the chosen parameters correspond to the global maximum. If another DNM routine identifies the same parameters from the same log-likelihood then more certainty can be given that the parameter values are related to the true global maximum.

The purpose of this analysis was to model CO_2 extremes and it required data in point process form. A threshold was selected using the mean residual life (MRL) plot and a rule of thumb. There is a large degree of subjectivity when deciding the value of the threshold and the level of the threshold greatly affects how many CO_2 events are considered extreme. A lower threshold for either the dry or wet season would result in more exceedances, which would change the

event history and corresponding CO_2 mark values and distribution. These changes may result in a different model formulation with different parameters as the Hawkes log-likelihood depends on the event history and the marked Hawkes log-likelihood depends on the event history and CO_2 marks. However, this is dependent on the seasonal dataset with which the threshold is found, as a slightly lower threshold may only result in a few more events being considered extreme.

Moreover, the choice of threshold was carefully selected as the objective of this project is to model extreme values which are the right-tail observations of the CO_2 distribution. As such, careful threshold selection was needed to isolate the right-tail of the distribution and to determine if the events are representative of what can truly be considered environmentally extreme. Hence, other methods should be employed in the future to more accurately determine the threshold above which CO_2 events are truly climatically extreme.

To account for the seasonality present in the CO_2 data, the dataset was split into dry season 1, wet season and dry season 2. This enabled varied seasonal thresholds to be determined, resulting in some events in each season being considered seasonally extreme. This separation of seasons does have limitations as it leads to the formulation of three separate models, resulting in the process being captured in subsections. These models are not able to be generalised to model the CO_2 level throughout the two years due to their seasonal specificity. An alternative approach is to include a seasonal spline function in the threshold determination process, enabling seasonal extremes to be captured by a single threshold with seasonal variation. This could capture the seasonality well as seen in Figure 4 and would enable the entire process to be captured in a single, more generally applicable model.

The issue with splitting the seasons is that there is often seasonal variation present within the year. The wet and dry seasons may not have a clear division as to where one season ends and the other begins as there is often fluctuation in weather patterns. This yearly variation could be accounted for by including covariates such as rainfall, air temperature, soil temperature and moisture. Covariates may be particularly useful in modelling environmental time series, due to the significant degree of interaction between variables often observed in environmental systems. These variables may capture the underlying process better than the marks, as they do not appear to improve the model fit.

10 Conclusion and final remarks

This paper mainly focused on the suitability of using a point process to model environmental time series, in particular, CO_2 extreme events where there is demonstrable temporal clustering present between the events. The proposed models consisted of a Poisson process, the Hawkes process and a more complex extension of the Hawkes process which included CO_2 magnitude marks of the extreme events. Before modelling, the two-year CO_2 observations were split into three complete seasons, two dry and one wet. Thereafter the extreme CO_2 readings were extracted by setting a season-dependent threshold, such that the CO_2 values exceeding the

threshold values are considered extreme.

It has been demonstrated that the Hawkes process model with an exponential decay function appears to be suitable for extreme CO_2 observations, especially in the wet seasons, when the extreme events are less severely clustered. When the extreme events are more tightly clustered with larger inter-arrival times between the clusters, as in both dry seasons 1 and 2, the Hawkes model appears less able to account for the extreme clustering and has a tendency to overestimate the inter-arrival times. Future considerations should be made with regard to the specific formulation of the decay function, as there is room for improvement of the Hawkes model and perhaps a different decay function may produce better estimations of event inter-arrival times. A possible alternative formulation would be to use the power law decay function.

Interestingly, the addition of CO_2 magnitude marks with the marked Hawkes model offered little to no improvement from the Hawkes model for any of the three seasons, indicating that the manner in which the marks were included in this model may not capture the underlying system and that the added complication thereof is unnecessary and provides no gain in model fit. Although this is not surprising given that, from an atmospheric standpoint, the magnitude of a CO_2 extreme event will most likely not excite another event. Other impact functions and mark distributions could be used to fit new models to determine if some of the conditional intensity may be attributed to marks or if marks play no role as expected.

Although the CO_2 magnitude marks do not play a role in this particular formulation of the environmental time series, that does not indicate that they would not account for some of the conditional intensity in other environmental systems consisting of extreme events. For instance, the magnitude of an earthquake is an important feature that can trigger another earthquake.

This analysis used a univariate point process to model environmental extremes over time because in the exploration of literature prior to conducting the analysis, CO_2 was found to not exhibit a mutually exciting relationship with other environmental factors. However, for other environmental time series, there may be a mutually exciting relationship through which the system may be able to be modelled more holistically rather than using marks.

Lastly, future considerations should be made with regards to modelling the extreme CO_2 events over time entire time interval rather than modelling the seasonal extremes to capture the system more holistically. As discussed above, this could be done through the inclusion of a covariate or a smoothed seasonal spline in the threshold-determining function.

11 Appendix

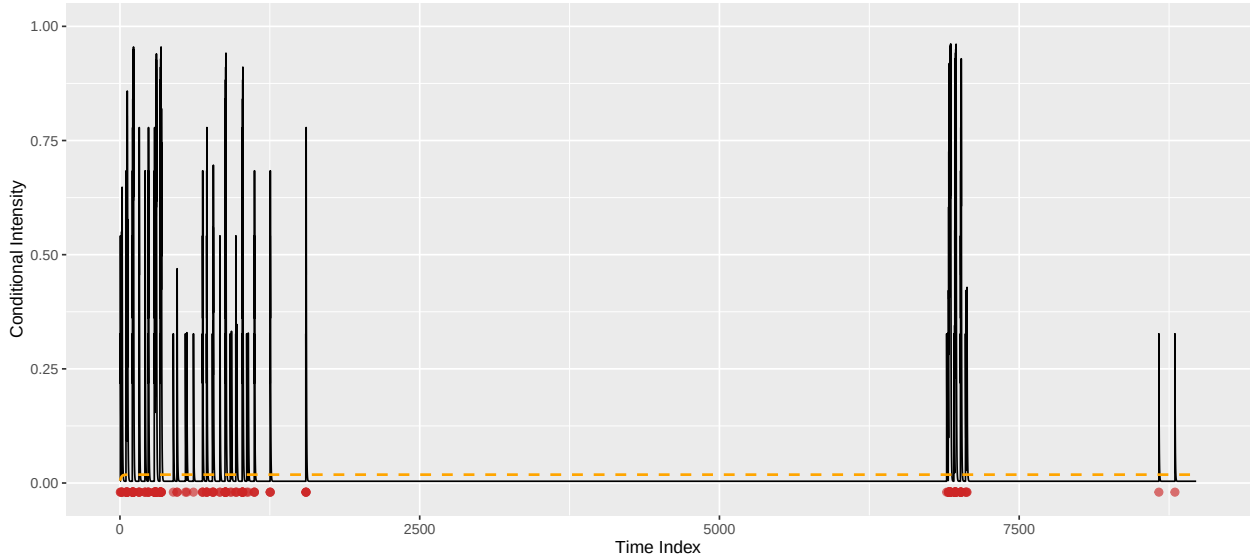

Figure 33: Dry 1 extreme Hawkes conditional intensity with parameters $\lambda = 0.0039$, $\alpha = 0.323$ and $\beta = 0.409$ given by the black line. Events are given by red dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

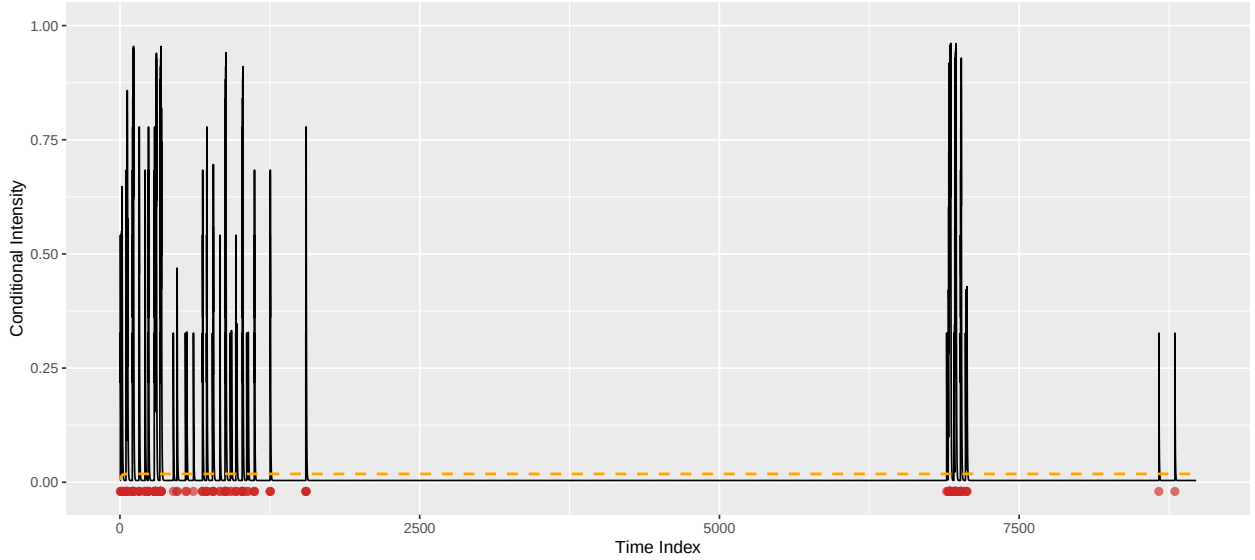

Figure 34: Dry 1 extreme marked Hawkes conditional intensity with parameters $\lambda = 0.0039$, $\alpha = 0.323$, $\beta = 0.409$ and $\delta = 0$ given by the black line. Events are given by red dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

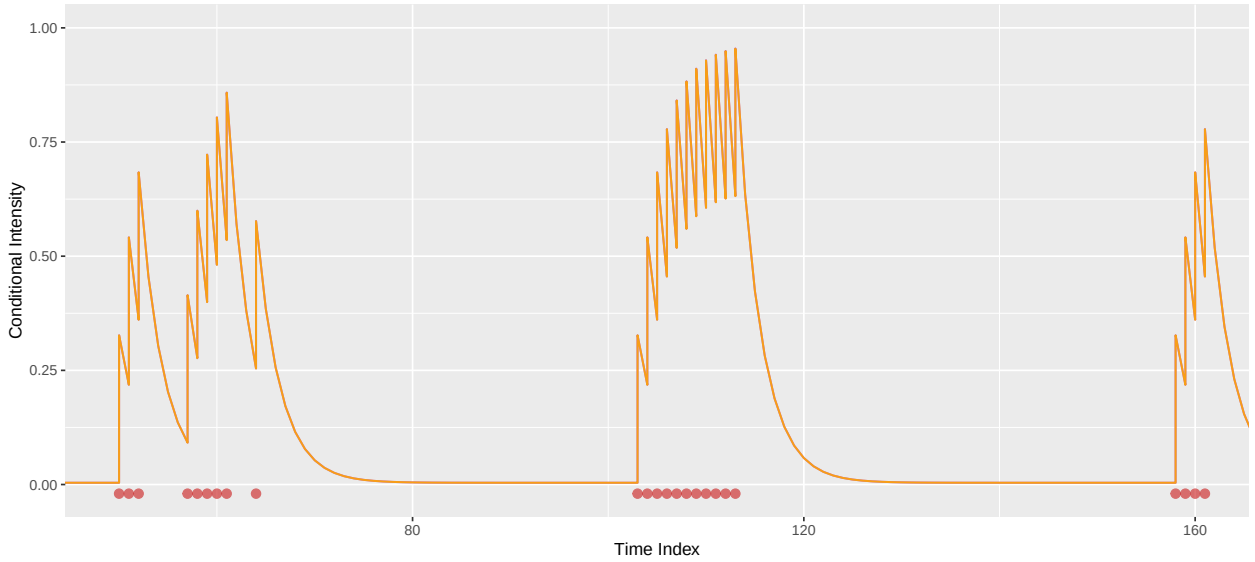

Figure 35: Subsection of Dry 1 extreme Hawkes conditional intensity with parameters $\lambda = 0.0039$, $\alpha = 0.3230$ and $\beta = 0.4093$ given by the purple line. Marked Hawkes conditional intensity with parameters $\lambda = 0.0039$, $\alpha = 0.323$, $\beta = 0.4093$ and $\delta = 0$ given by the black line. Events are given by red dots. The orange line is overlaid on the purple and therefore the purple cannot be seen.

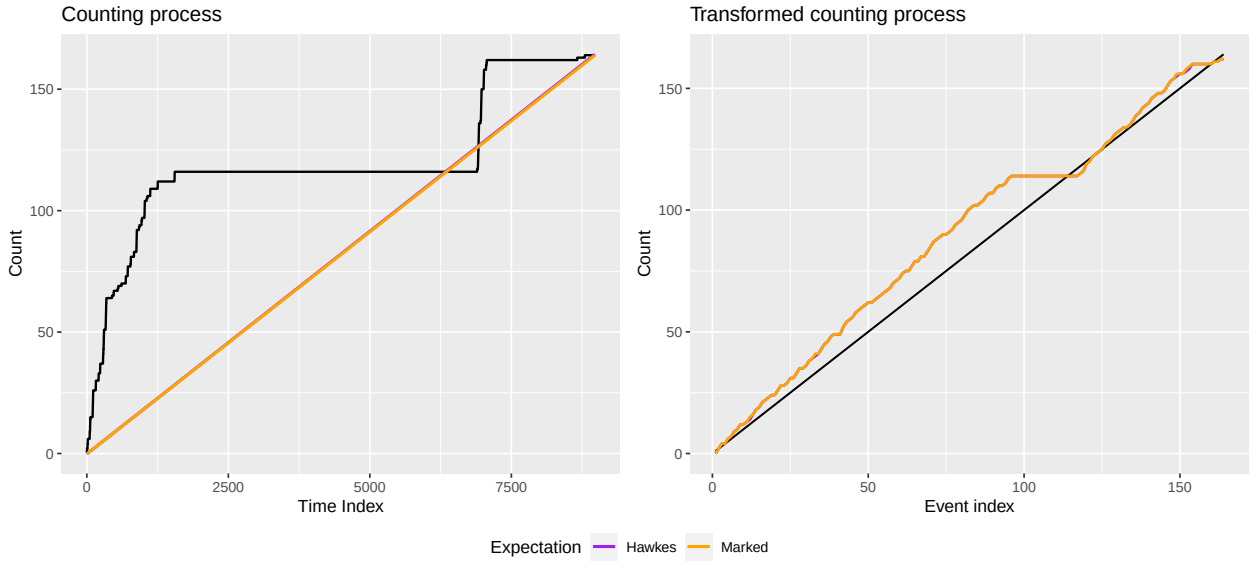

Figure 36: Fit checking plots comparison for the Hawkes and marked Hawkes process model for dry extreme 1. (A) The counting process (black line) with the expectation of the counting process, denoted $\mathbb{E}[N(t)]$. (B) The counting process (black line) resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the transformed counting process, denoted $\mathbb{E}[N^*(t)]$. The orange line is overlaid on the purple and therefore the purple cannot be seen well.

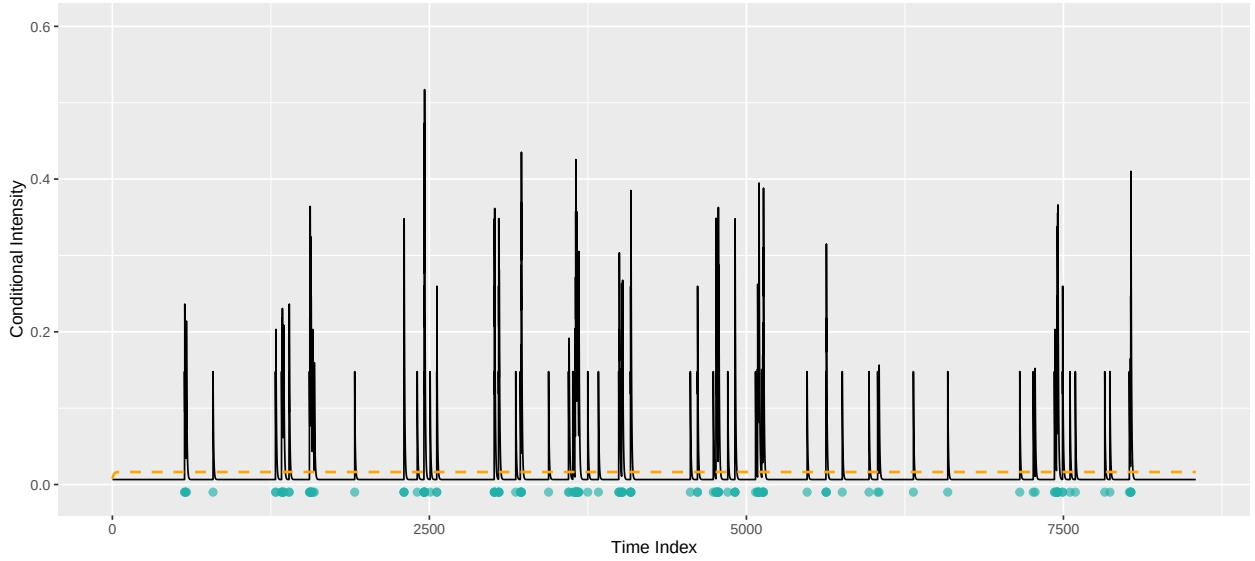

Figure 37: Wet extreme Hawkes conditional intensity with parameters $\lambda = 0.0065$, $\alpha = 0.141$ and $\beta = 0.235$ given by the black line. Events are given by blue dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

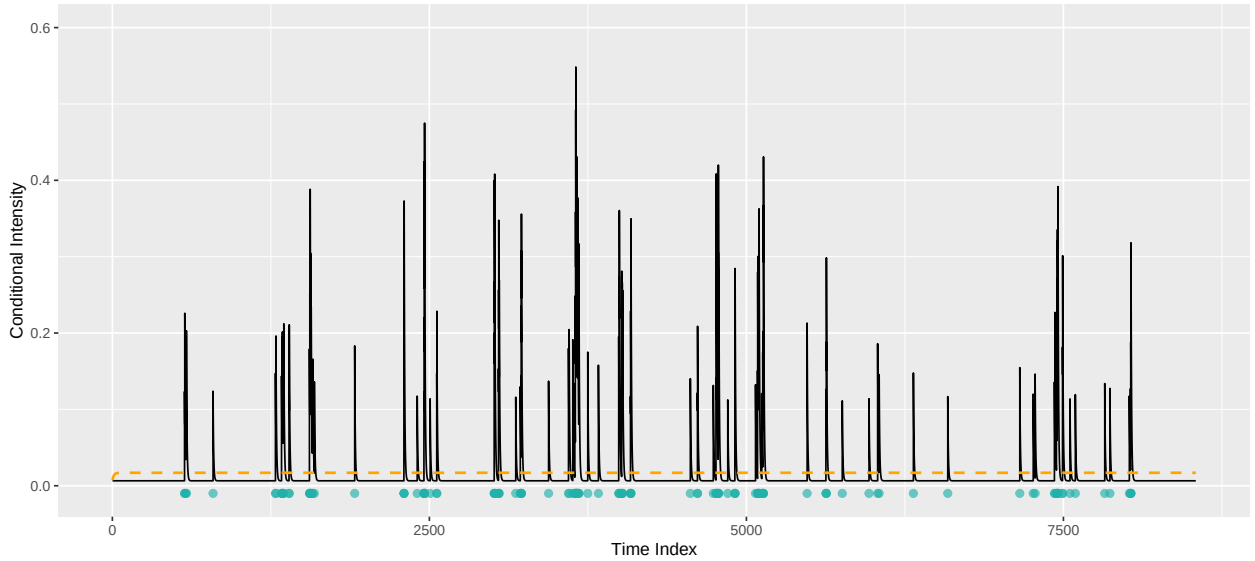

Figure 38: Wet extreme marked Hawkes conditional intensity with parameters $\lambda = 0.0065$, $\alpha = 0.105$, $\beta = 0.230$ and $\delta = 0.0098$ given by the black line. Events are given by blue dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

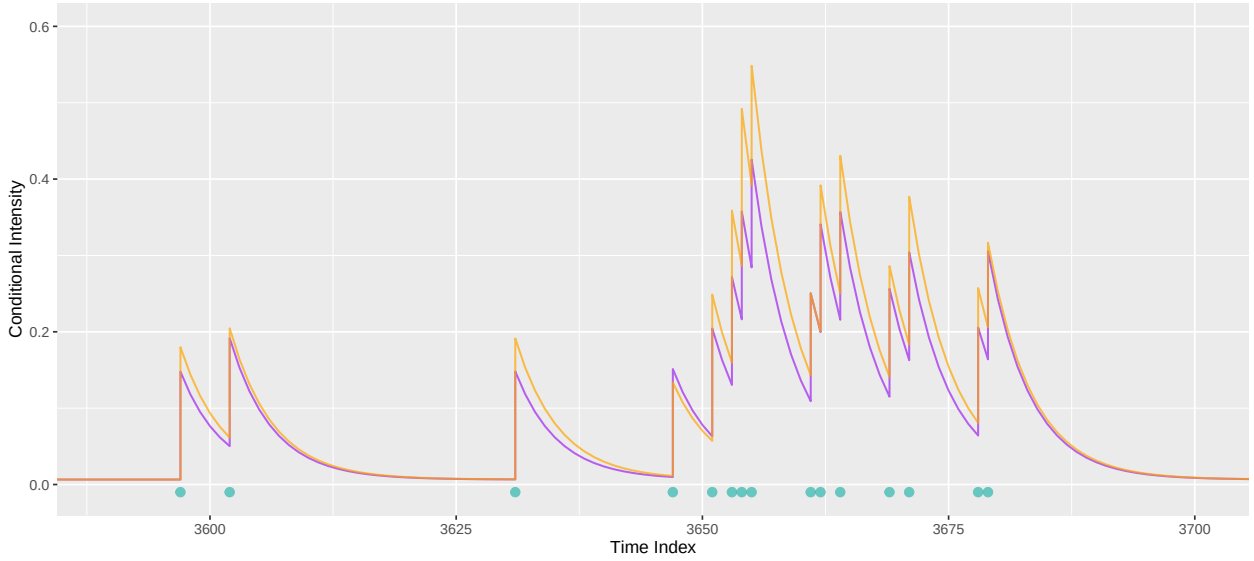

Figure 39: Subsection of wet extreme Hawkes conditional intensity with parameters $\lambda = 0.0039$, $\alpha = 0.323$ and $\beta = 0.409$ given by the purple line. Marked Hawkes conditional intensity with parameters $\lambda = 0.0065$, $\alpha = 0.105$, $\beta = 0.230$ and $\delta = 0.0098$ given by the orange line. Events are given by blue dots.

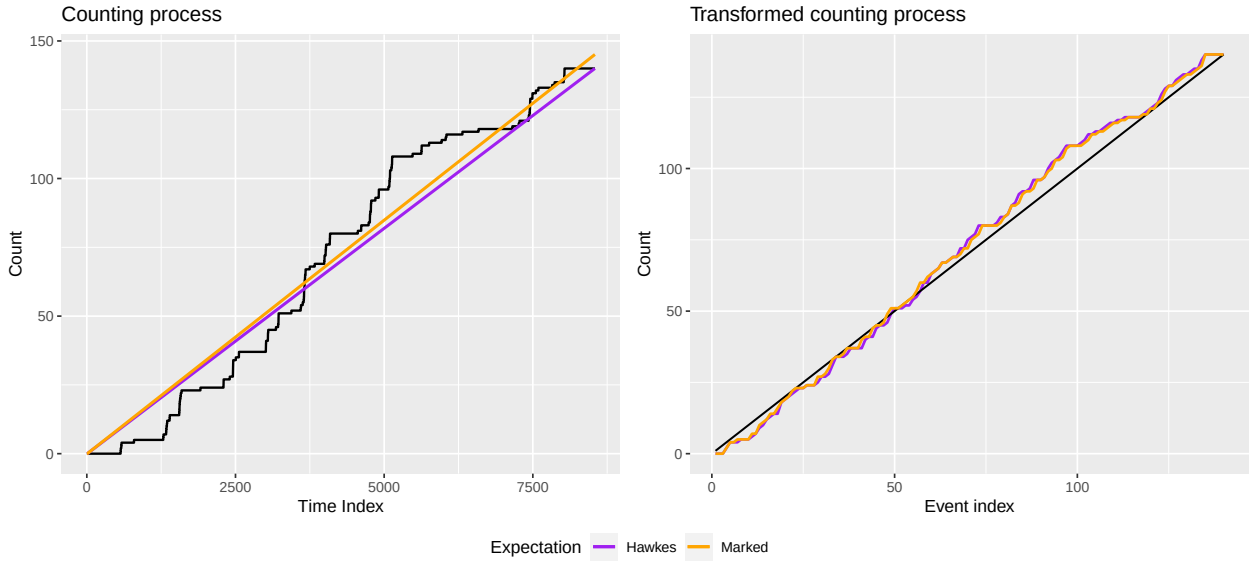

Figure 40: Fit checking plots comparison for the Hawkes and marked Hawkes process model for wet extreme. (A) The counting process (black line) with the expectation of the counting process, denoted $\mathbb{E}[N(t)]$. (B) The counting process (black line) resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the transformed counting process, denoted $\mathbb{E}[N^*(t)]$.

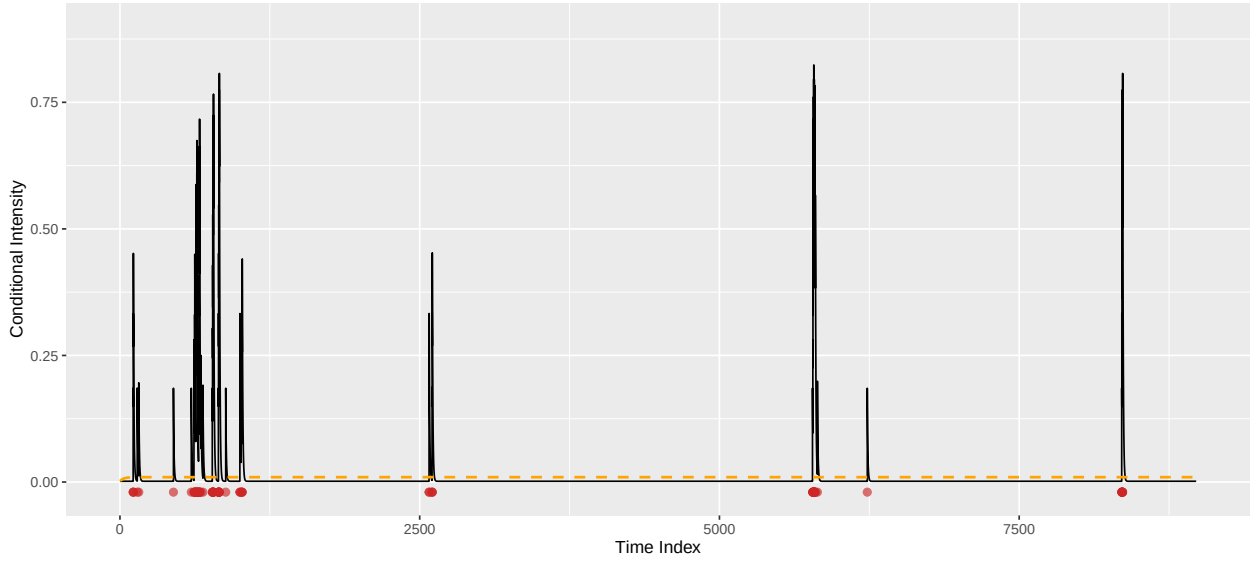

Figure 41: Dry 2 extreme Hawkes conditional intensity with parameters $\lambda = 0.0015$, $\alpha = 0.184$ and $\beta = 0.218$ given by the black line. Events are given by red dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

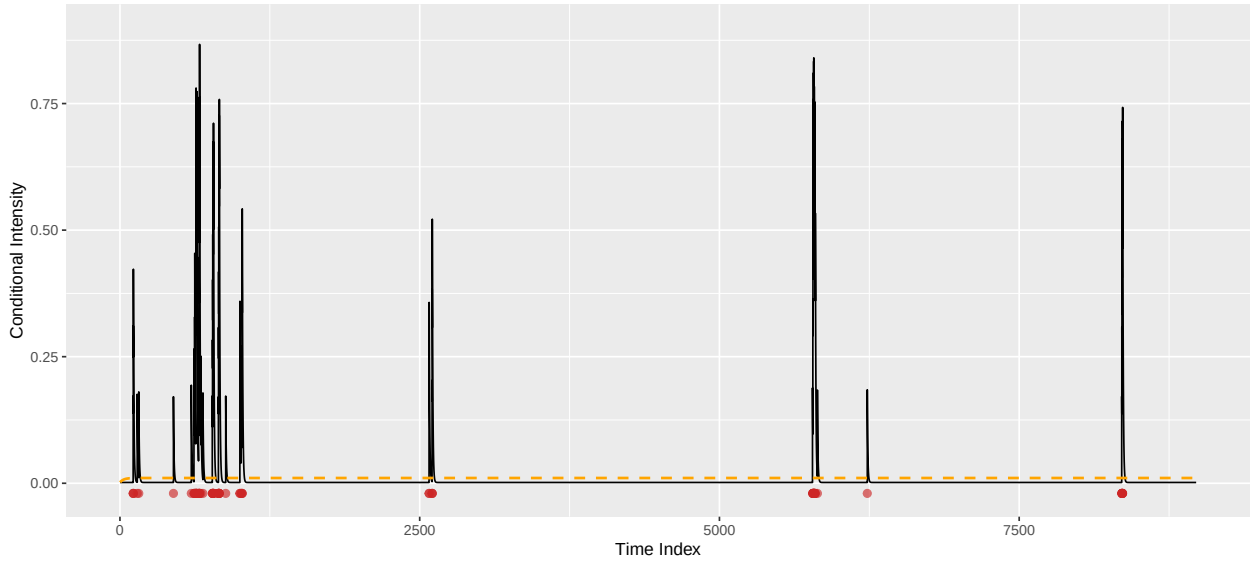

Figure 42: Dry 2 extreme Marked Hawkes conditional intensity with parameters $\lambda = 0.0016$, $\alpha = 0.169$, $\beta = 0.223$ and $\delta = 0.0046$ given by the black line. Events are given by red dots and the orange dotted line is the expectation of the intensity, $\mathbb{E}[\lambda(t)]$.

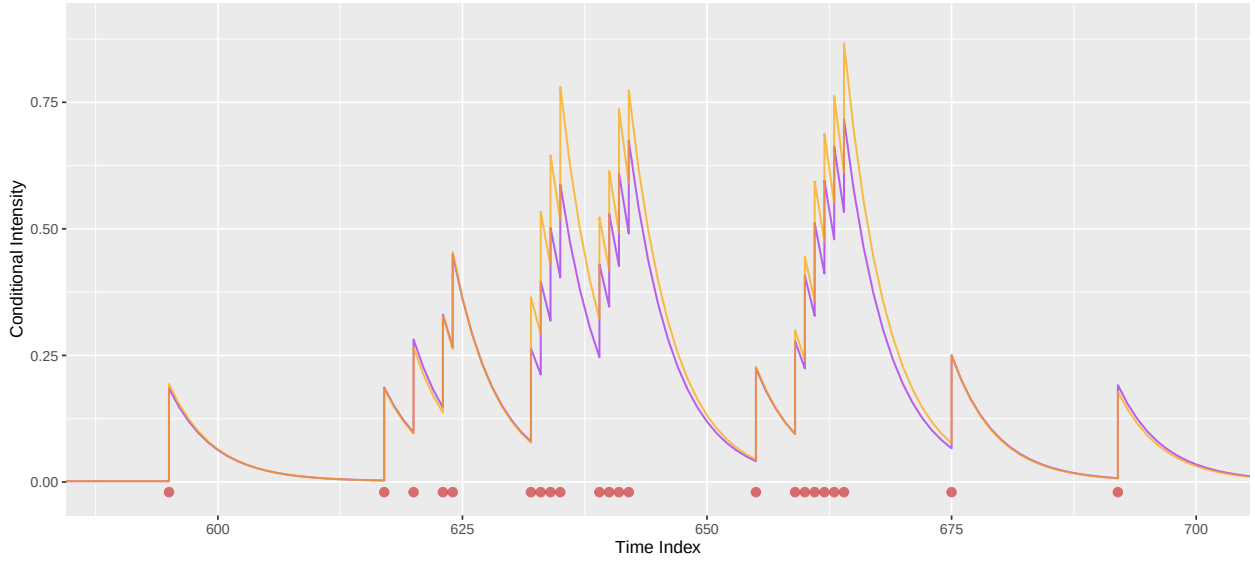

Figure 43: Subsection of dry 2 extreme Hawkes conditional intensity with parameters $\lambda = 0.0015$, $\alpha = 0.184$ and $\beta = 0.218$ given by the purple line. Marked Hawkes conditional intensity with parameters $\lambda = 0.0016$, $\alpha = 0.169$, $\beta = 0.223$ and $\delta = 0.0046$ given by the orange line. Events are given by red dots.

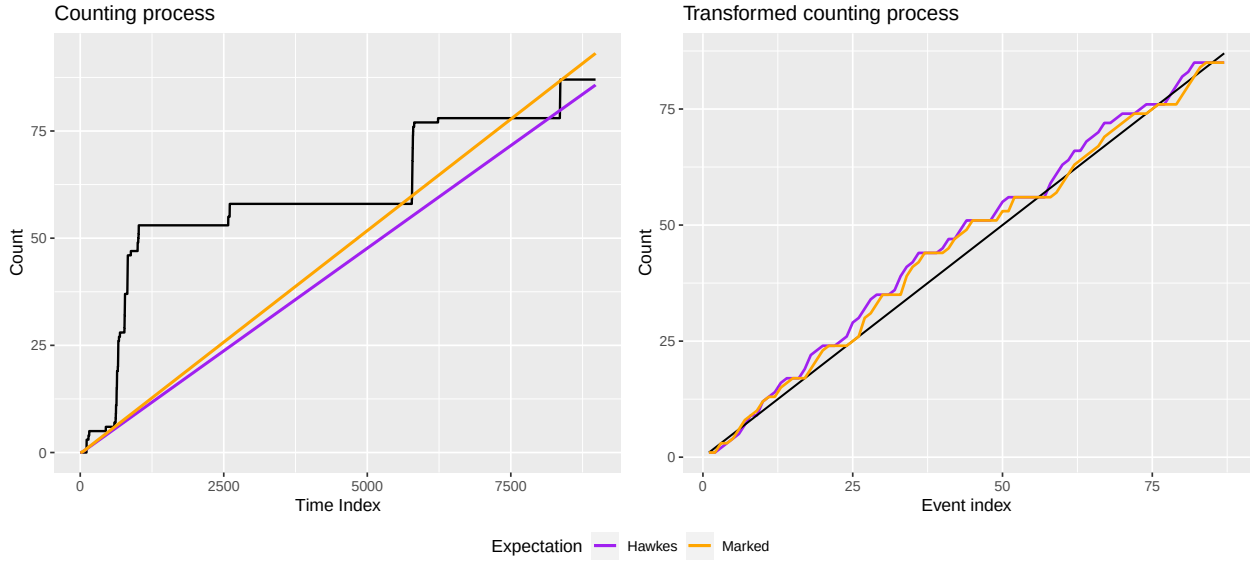

Figure 44: Fit checking plots comparison for the Hawkes and marked Hawkes process model for dry extreme 2. (A) The counting process (black line) with the expectation of the counting process, denoted $\mathbb{E}[N(t)]$. (B) The counting process (black line) resulting from the random time change transformation of the data ($N^*(t)$ where $t_i^* = \Lambda(t_i)$) with the expectation of the transformed counting process, denoted $\mathbb{E}[N^*(t)]$.

12 References

Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), pp.716-723.

Al-Bayati, R.M., Adeeb, H.Q., Al-Salihi, A.M. and Al-Timimi, Y.K., 2020, December. The relationship between the concentration of carbon dioxide and wind using GIS. In AIP Conference Proceedings (Vol. 2290, No. 1, p. 050042). AIP Publishing LLC.

Balderama, E., Schoenberg, F.P., Murray, E. and Rundel, P.W., 2012. Application of branching models in the study of invasive species. Journal of the American Statistical Association, 107(498), pp.467-476.

Bartlett, M.S., 1963a. The spectral analysis of point processes. Journal of the Royal Statistical Society: Series B (Methodological), 25(2), pp.264-281.

Bartlett, M.S., 1963b. Statistical estimation of density functions. Sankhyā: The Indian Journal of Statistics, Series A, pp.245-254.

Boano-Danquah, J., Sigauke, C. and Kyei, K.A., 2020. Analysis of extreme peak loads using point processes: An application using South African data. IEEE Access, 8, pp.146105-146115.

Bommier, E., 2014. Peaks-over-threshold modelling of environmental data.

Bonnet, A., Herrera, M.M. and Sangnier, M., 2021. Maximum Likelihood Estimation for Hawkes Processes with self-excitation or inhibition. Statistics and Probability Letters, 179, p.109214.

Campbell Scientific, 2022a. IRGASON: A New Eddy Covariance Analyzer. Available at: https://www.campbellsci.asia/new-irgason [2022, 1 October]

Campbell Scientific, 2022b. Expandable Closed-Path Eddy-Covariance System with EC155 and Pump Module. Available at: https://www.campbellsci.com.au/cpec306 [2022, 27 July]

Chiang, W.H., Liu, X. and Mohler, G., 2022. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. International journal of forecasting, 38(2), pp.505-520.

Coles, S.G., 2001. An Introduction to Statistical Modelling of Extreme Values, Springer, London.

Cox, D.R. and Isham, V., 1980. Point processes (Vol. 12). CRC Press.

Cox, D.R. and Lewis, P.A., 1966. The Statistical Analysis of Series of Events.

Cui, L., Hawkes, A. and Yi, H., 2020. An elementary derivation of moments of Hawkes processes. Advances in Applied Probability, 52(1), pp.102-137.

Dalelane, C. and Deutschländer, T., 2013. A robust estimator for the intensity of the Poisson point process of extreme weather events. Weather and Climate Extremes, 1, pp.69-76.

Daley, D.J. and Vere-Jones, D., 2003. An introduction to the theory of point processes: volume I: elementary theory and methods. Springer New York.

Dassios, A. and Zhao, H., 2013. Exact simulation of Hawkes process with exponentially decaying intensity. Electronic Communications in Probability, 18, pp.1-13.

Descombes, X. and Zerubia, J., 2002. Marked point process in image analysis. IEEE Signal Processing Magazine, 19(5), pp.77-84.

Diebold, F.X., Schuermann, T. and Stroughair, J.D., 1998. Pitfalls and opportunities in the use of extreme value theory in risk management. In Decision technologies for computational finance (pp. 3-12). Springer, Boston, MA.

DuMouchel, W.H., 1983. Estimating the stable index α in order to measure tail thickness: A critique. the Annals of Statistics, 11(4), pp.1019-1031.

Embrechts, P., Liniger, T. and Lin, L., 2011. Multivariate Hawkes processes: an application to financial data. Journal of Applied Probability, 48(A), pp.367-378.

Ferreira, A., de Haan, L. and Peng, L., 2003. On optimising the estimation of high quantiles of a probability distribution, Statistics, 37, 401–434.

Fox, E.W., Schoenberg, F.P. and Gordon, J.S., 2016. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. The Annals of Applied Statistics, 10(3), pp.1725-1756.

Fuest, A., 2009. Modelling temporal dependencies of extreme events via point processes (Doctoral dissertation, Institut für Statistik).

Garetto, M., Leonardi, E. and Torrisi, G.L., 2021. A time-modulated Hawkes process to model the spread of COVID-19 and the impact of countermeasures. Annual reviews in control, 51, pp.551-563.

Gavrikov, V. and Stoyan, D., 1995. The use of marked point processes in ecological and environmental forest studies. Environmental and ecological statistics, 2(4), pp.331-344.

Gebbie, T., 2022. STA3045F: Advanced Stochastic Processes, Part II: Modelling Random Events. Department of Statistical Sciences, University of Cape Town, v.1.2.

Gilli, M. and Kellezi, E., 2006. An application of extreme value theory for measuring financial risk. Computational Economics, 27(2), pp.207-228.

Grothe, O., Korniichuk, V. and Manner, H., 2014. Modelling multivariate extreme events using self-exciting point processes. Journal of Econometrics, 182(2), pp.269-289.

Guo, C. and Luk, W., 2013, September. Accelerating maximum likelihood estimation for Hawkes point processes. In 2013 23rd International Conference on Field Programmable Logic and Applications (pp. 1-6). IEEE.

Gupta, A., Farajtabar, M., Dilkina, B. and Zha, H., 2018, January. Discrete Interventions in Hawkes Processes with Applications in Invasive Species Management. In IJCAI (pp. 3385-3392).

Hardiman, S.J., Bercot, N. and Bouchaud, J.P., 2013. Critical reflexivity in financial markets: a Hawkes process analysis. The European Physical Journal B, 86(10), pp.1-9.

Hawkes, A.G., 1971. Point spectra of some mutually exciting point processes. Journal of the Royal Statistical Society: Series B (Methodological), 33(3), pp.438-443.

Hawkes, A.G., 2018. Hawkes processes and their applications to finance: a review. Quantitative Finance, 18(2), pp.193-198.

Hawkes, A.G. and Oakes, D., 1974. A cluster process representation of a self-exciting process. Journal of Applied Probability, 11(3), pp.493-503.

Holden, L., Sannan, S. and Bungum, H., 2003. A stochastic marked point process model for earthquakes. Natural Hazards and Earth System Sciences, 3(1/2), pp.95-101.

Hovenden, M.J., Newton, P.C. and Wills, K.E., 2014. Seasonal not annual rainfall determines grassland biomass response to carbon dioxide. Nature, 511(7511), pp.583-586.

Hu, Y. and Scarrott, C., 2018. "evmix: An R package for Extreme Value Mixture Modeling, Threshold Estimation and Boundary Corrected Kernel Density Estimation." Journal of Statistical Software, *84*(5), 1-27. doi:10.18637/jss.v084.i05 https://doi.org/10.18637/jss.v084.i05

Kim, J., Verma, S.B. and Clement, R.J., 1992. Carbon dioxide budget in a temperate grassland ecosystem. Journal of Geophysical Research: Atmospheres, 97(D5), pp.6057-6063.

Kobayashi, R. and Lambiotte, R., 2016, March. TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. In Tenth International AAAI Conference on Web and Social Media.

Koh, J., Pimont, F., Dupuy, J.L. and Opitz, T., 2021. Spatiotemporal wildfire modelling through point processes with moderate and extreme marks. arXiv preprint arXiv:2105.08004.

Lapham, B.M., 2014. Hawkes processes and some financial applications (Master's thesis, University of Cape Town).

Laub, P.J., Lee, Y. and Taimre, T., 2021. The Elements of Hawkes Processes.

Li, L. and Zha, H., 2015, February. Energy usage behavior modeling in energy disaggregation via marked Hawkes process. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

Liang, S. and Wang, J. eds., 2019. Advanced remote sensing: terrestrial information extraction and applications. Academic Press.

Liniger, T.J., 2009. Multivariate Hawkes processes (Doctoral dissertation, ETH Zurich).

Loretan, M. Philips, P.C.B., 1994. Testing the covariance stationarity of heavy tailed time series: an overview of the theory with applications to several financial datasets, J. R. Statist. Soc. D, 1, 211–248.

Lukasik, M., Srijith, P.K., Vu, D., Bontcheva, K., Zubiaga, A. and Cohn, T., 2016, August. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 393-398).

Malaviya, J., 2021. Survey on modelling intensity function of Hawkes process using neural models. arXiv preprint arXiv:2104.11092.

McNeil, A.J., Frey, R. and Embrechts, P., 2005. Quantitative risk management: concepts, techniques and tools-revised edition. Princeton university press.

Mei, H. and Eisner, J.M., 2017. The neural Hawkes process: A neurally self-modulating multivariate point process. Advances in neural information processing systems, 30.

Morales, F.E.C. and Rodrigues, D.T., 2022. Spatiotemporal nonhomogeneous Poisson model with a seasonal component applied to the analysis of extreme rainfall. Journal of Applied Statistics, pp.1-19.

Ngailo, T., Shaban, N., Reuder, J., Rutalebwa, E. and Mugume, I., 2016. Non homogeneous Poisson process modelling of seasonal extreme rainfall events in Tanzania. International journal of science and research (IJSR), 5(10), pp.1858-1868.

Nie, H.R., Zhang, X., Li, M., Dolgun, A. and Baglin, J., 2020, October. Modelling user influence and rumor propagation on Twitter using Hawkes processes. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 637-656). IEEE.

Oakes, D., 1975. The Markovian self-exciting process. Journal of Applied Probability, 12(1), pp.69-77.

Ogata, Y., 1981. On Lewis' simulation method for point processes. IEEE transactions on information theory, 27(1), pp.23-31.

Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association, 83(401), pp.9-27.

Ogata, Y., 1998. Space-time point-process models for earthquake occurrences. Annals of the Institute of Statistical Mathematics, 50(2), pp.379-402.

Ogata, Y. and Zhuang, J., 2006. Space-time ETAS models and an improved extension. Tectonophysics, 413(1-2), pp.13-23.

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Räsänen, M., Aurela, M., Vakkari, V., Beukes, J.P., Tuovinen, J.P., Van Zyl, P.G., Josipovic, M., Venter, A.D., et al., 2017. Carbon balance of a grazed savanna grassland ecosystem in South Africa. Biogeosciences, 14(5), pp.1039-1054.

Reiss, R.D. and Thomas, M., 1997. Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhauser Verlag, Basel.

Ribatet, M. and Dutang, C., 2022. POT: Generalized Pareto Distribution and Peaks Over Threshold. R package version 1.1-10.

Scarrott, C. and MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT-Statistical journal, 10(1), pp.33-60.

Smith, J.A., 1993. Marked point process models of raindrop-size distributions. Journal of Applied Meteorology and Climatology, 32(2), pp.284-296.

Smith, R.L., 1989. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. Statistical Science, pp.367-377.

South African Environmental Observation Network (SAEON), 2022. KIMTRI metadata [Dataset].

Zepp, R.G., Miller, W.L., Burke, R.A., Parsons, D.A. and Scholes, M.C., 1996. Effects of moisture and burning on soil atmosphere exchange of trace carbon gases in a southern African savanna. Journal of Geophysical Research: Atmospheres, 101(D19), pp.23699-23706.

Veen, A. and Schoenberg, F.P., 2008. Estimation of space–time branching process models in seismology using an em–type algorithm. Journal of the American Statistical Association, 103(482), pp.614-624.

Wang, T., Bebbington, M. and Harte, D., 2012. Markov-modulated Hawkes process with stepwise decay. Annals of the Institute of Statistical Mathematics, 64(3), pp.521-544.

Weather Spark, n.d. Kimberley Climate, Weather By Month, Average Temperature (South Africa) - Weather Spark. Available at: https://weatherspark.com/y/90479/Average-Weather-in-Kimberley-South-Africa-Year-Round#Sections-Precipitation [2022, 4 September]

Williams, C.A., Hanan, N., Scholes, R.J. and Kutsch, W., 2009. Complexity in water and carbon dioxide fluxes following rain pulses in an African savanna. Oecologia, 161(3), pp.469-480.

Xiao, S., Yan, J., Yang, X., Zha, H. and Chu, S., 2017, February. Modelling the intensity function of the point process via recurrent neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

Zhang, Q., Lipani, A., Kirnap, O. and Yilmaz, E., 2020, November. Self-attentive Hawkes process. In International conference on machine learning (pp. 11183-11193). PMLR.

Zucchini, W. and MacDonald, I.L., 2009. Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC.