UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO de Engenharia Geográfica, Geofísica e Energia

Origin of power-law behaviour in the size distribution of extreme events of gross primary productivity

Elmar Geerlings

Mestrado em Ciências Geofísicas
Meteorologia e Oceanografia

Dissertação orientada por:
Dr. Pedro Manuel Alberto de Miranda
Dr. Christian Reick

2021

Abstract

A quite interesting find by Zscheischler et al. in 2013 [1] was that the size distribution of extreme events in observation data of gross primary productivity (GPP) follows a power-law in the form $p(x) \sim x^{-\alpha}$. This power-law holds for different regions in the world with similar values for the scaling parameter α.

The goal of this thesis is to unravel the origin of this power-law behaviour. This behaviour might originate from the GPP distribution itself, or perhaps have a more mathematical origin. Thus, the main research question to be answered in this study is: ”What is the origin of the power-law behaviour in the size distribution of GPP extreme events?”

With data from a control simulation from CMIP6 (Coupled Model Intercomparison Project Phase 6), I used the methodology from Zscheischler et al. for finding extreme events in simulation data for GPP. The power-law is not found in the distribution of GPP itself, thus its origins are sought in the clustering mechanisms behind the extreme event analysis. Percolation theory is hypothesised as an explanation behind the power-law behaviour, based on the fact that both GPP extremes and percolation theory are concerned with clusters made out of a certain fraction of the data. This certain fraction is made up by ”percentiles” for GPP extremes and ”probability” in percolation theory. The exponent α for the power-law in the size distribution of GPP is related to the exponent τ describing cluster sizes in percolation theory by the relation $\tau = \alpha + 1$. However, there are some differences in the power-law scaling behaviour between GPP extremes and percolation theory, namely concerning the difference in the value of the voxels (i.e. 3D pixels) of GPP, correlations in time and space, and the restriction of GPP values to land. The GPP data is altered step by step to eliminate these differences to make the data more akin to the situation of percolation theory, which assumes uncorrelated data. This is done by considering cluster sizes instead of event sizes, randomizing the data by ”shuffling” and using synthetic datasets, producing results of power-law scaling behaviour that are closer to percolation theory. The most rigorous shuffled data and the synthetic data had power-law scaling behaviour that was especially close to percolation theory. Based on this, it can be said that the clustering mechanisms behind extreme event analysis are similar to the clustering in percolation theory and that therefore percolation theory can be considered as a reasonable explanation behind the power-law in GPP extremes. The size distribution of precipitation, sensible heat and latent heat also display power-law behaviour similar to GPP, indicating that this power-law is not exclusive to GPP. All in all it can be concluded that the origin of the power-law behaviour does not depend on GPP, in general it does not depend on the data itself but on the clustering mechanisms underlying percolation theory.

Keywords: gross primary productivity (GPP), extreme events, power-law, percolation theory

Sumário

Uma descoberta bastante interessante de Zscheischler et al. em 2013 [1] foi de que a distribuição do tamanho de eventos extremos em dados de observação de produtividade primária bruta (GPP, do inglês gross primary productivity) segue uma lei de potência na forma $p(x) \sim x^{-\alpha}$. Tal lei de potência é válida em diferentes regiões do mundo, com valores semelhantes para o parâmetro de escala α.

O objectivo desta tese é revelar a origem deste comportamento de lei de potência. Este comportamento pode originar-se da própria distribuição de GPP ou talvez ter uma origem mais matemática. Assim, a principal questão de investigação a ser respondida neste estudo é: "Qual é a origem do comportamento de lei de potência da distribuição do tamanho de eventos extremos na GPP"?

Com dados de uma simulação de controle do CMIP6 (Coupled Model Intercomparison Project Phase 6), utilizei a metodologia de Zscheischler et al. para encontrar eventos extremos nos dados de simulação de GPP. A lei de potência não é encontrada na distribuição da própria GPP, pelo que as suas origens são procuradas nos mecanismos de aglomeração por detrás da análise de eventos extremos. A teoria da percolação é colocada como hipótese para explicar o comportamento de lei de potência, com base no facto de que tanto os extremos de GPP quanto a teoria da percolação estão relacionados a aglomerados compostos a partir de uma certa fracção dos dados. Esta certa fracção é constituída por "percentis" no caso dos extremos de GPP e por "probabilidade" no caso da teoria da percolação. O expoente α da lei de potência na distribuição de tamanho de GPP está associado ao expoente τ, que descreve tamanhos de aglomerados na teoria da percolação, pela relação $\tau = \alpha + 1$. Contudo, existem algumas diferenças no comportamento de lei de potência entre os extremos de GPP e a teoria da percolação, nomeadamente no que diz respeito à diferença no valor dos voxels (i.e. pixels em 3D) de GPP, correlações no tempo e espaço, e à restrição dos valores de GPP aos continentes. Os dados de GPP são alterados passo a passo para eliminar estas diferenças de modo a torná-los mais semelhantes à situação da teoria da percolação, que assume dados não correlacionados. Isto é feito considerando os tamanhos dos aglomerados ao invés dos tamanhos dos eventos, randomizando os dados através de um processo de "embaralhamento", e utilizando conjuntos de dados sintéticos, produzindo resultados de comportamentos de lei de potência que estão mais próximos à teoria da percolação. Os dados embaralhados mais rigorosamente e os dados sintéticos apresentaram um comportamento de lei de potência especialmente próximo daquele na teoria da percolação. Com base nisto, pode-se dizer que os mecanismos de aglomeração por detrás da análise de eventos extremos são semelhantes à aglomeração na teoria da percolação e que, portanto, a teoria da percolação pode ser considerada como uma explicação razoável por detrás da lei de potência nos extremos de GPP. As distribuições de tamanho de precipitação, calor sensível e calor latente também apresentam um comportamento de lei de potência semelhante ao da GPP, indicando que esta lei de potência não ocorre unicamente para a GPP. Em suma, pode-se concluir que a origem do comportamento de lei de potência não depende da GPP, em geral não depende dos dados em si, mas sim dos mecanismos de aglomeração subjacentes à teoria da percolação.

Palavra-chave: produtividade primária bruta (GPP), eventos extremos, lei de potência, teoria da percolação

Contents

List of Figures

List of Tables

List of abbreviations and acronyms

GPP Gross Primary productivity
CO2 Carbon Dioxide
C Carbon
FAPAR Fraction of Absorbed Photosynthetically Active Radiation
EWE Extreme Weather Events
USA United States of America
2D 2 Dimensions
3D 3 Dimensions
CMIP6 Coupled Model Intercomparison Project Phase 6
MPI-ESM Max-Planck-Institue Earth System Model
neg Negative extremes
pos Positive extremes
GPP6000 GPP data containing 6000 timesteps
GPP600 GPP data containing 600 timesteps

List of symbols

p Probability, percentile
x Size of an extreme event
α Scaling parameter
p_c Percolation threshold
τ Powerlaw exponent
n_s Number of clusters of size s
s Cluster size
s_ζ Characteristic cluster size
σ Exponent describing the divergence from the percolation threshold
con Connectivity
y_s Cumulated value of the number of clusters larger than x

Chapter 1

Introduction and background

1.1 Introduction

Gross primary productivity (GPP) is a quantity that describes the exchange of CO2 between the biosphere and the atmosphere. More precisely, GPP can be defined as the gross carbon uptake by the terrestrial vegetation via photosynthesis [5]. A quite interesting find by Zscheischler et al.(2013)[1] was that extreme events in GPP follow a power-law in the form $p(x) \sim x^{-\alpha}$, where "extreme" is roughly defined as "highly anomalous". Moreover, this power-law holds for different regions in the world with similar values for the scaling parameter α [8].

In this study of Zscheischler et al. it has however not been investigated where this power-law behaviour in GPP extremes comes from. If GPP sizes would be a result of dynamics producing completely independent random numbers, one would get a Gaussian distribution with an exponential tail. Thus, the found power-law behaviour is an indication of some special mechanisms related with the dynamics of extremes. It is quite curious why this power-law occurs; this serves as the motivation for this master's thesis and is what I will investigate in it.

In another study, by Reichstein et al. (2013)[9], the aforementioned GPP extremes are linked to extreme weather events. Extreme weather can be defined as weather that lies outside a locale's normal range of weather intensity, as what is considered "extreme" at one location does not have to necessarily be considered extreme somewhere else [10]. There are several examples of where extreme weather events have an effect on ecosystems and their carbon balance by changing the net ecosystem carbon dioxide flux: A sustained decrease in net carbon uptake can shift forest ecosystems from a net carbon sink to a net carbon source [11]. Droughts can have a big impact on the mortality of vegetation, for example causing a large part of a dominant tree species to die in North America [12]. Other than droughts and heatwaves, extreme precipitation, extreme low temperatures and storms also have effects on the carbon cycle.

The topic and main research question of this master's thesis is as follows: "What is the origin of the power-law behaviour in the size distribution of GPP extreme events?". This origin may very well lie in the characteristics of GPP itself. Where GPP is influenced by extreme weather events, particularly droughts, heatwaves, extreme precipitation, extreme low temperatures and storms. Another explanation could be in the mechanisms behind the generation of power-laws such as critical behaviour caused by percolation [13].

1.2 Power-law behaviour of GPP extremes

The power-law that was found in the size distribution of GPP extreme events by Zscheischler et al. (2013) [1] serves as the main motivation for this master's thesis. I will look at the methodology and results of this study where this power-law has been found.

The acquisition of GPP data by Zscheischler et al. is done using a global dataset of the fraction of absorbed photosynthetically active radiation (FAPAR) [14]. The GPP data has a spatial resolution of 0.5° and spans from 1982-2011. The trend and seasonality is subtracted to be able to compare values across different seasons and years. The resulting values describe the GPP anomalies.

Zscheischer et al. defined extreme events by a set of percentiles; the 1st, 5th and 10th percentile. This means that extremes are detected that occur less than 1%, 5% or 10% of the time. Clusters of extremes, in space and time, are searched out by looking at connected components of voxels (i.e. 3D pixels) that contain an extreme, i.e. a value that is below a certain percentile. These clusters are then called 'extreme events'. The size distribution of these extreme events seems to follow a power-law and obeys the equation:

$$p(x) \sim x^{-\alpha}, \tag{1.1}$$

with some scaling parameter or exponent α, which is between 1 and 2 for most of the used data and percentiles. The size distribution of events in 5th-percentile GPP extremes can be seen in figure 1.1. It appears to follow a power-law for roughly two orders of magnitude, from 10^5 g C to 10^7 g C with scaling parameter $\alpha = 0.75$.

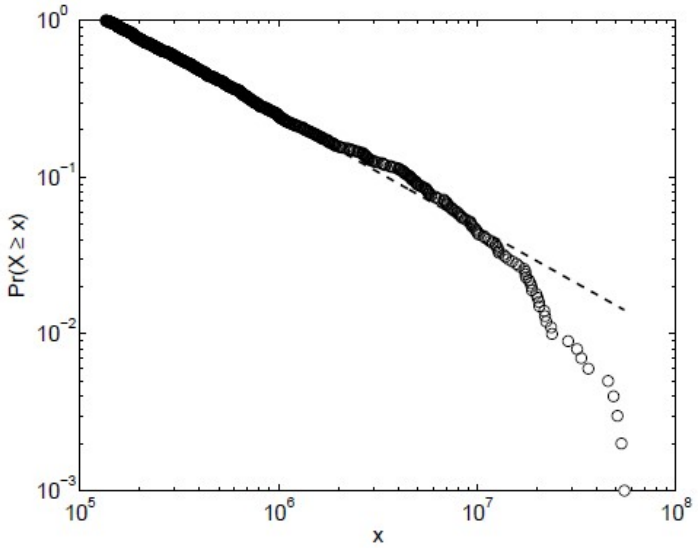

Figure 1.1: Power-law in the size distribution of events in 5th-percentile GPP extremes. x denotes the size of an individual event with its corresponding decrease in g C. Each circle is one event. The dashed line denotes the exact power-law distribution. Reproduced from [1]

In a subsequent study by Zscheischler et al. (2014) [2] these GPP extremes are further analyzed. Zscheischler et al. found here that a small amount of the biggest events dictate the global impact of all the extreme events. As can be seen in Figure 1.2, the biggest extremes largely determine the global GPP anomaly. The biggest 200 extreme events also only occur on only a small part (7%) of the spatiotemporal domain, revealing a strong spatial heterogeneity.

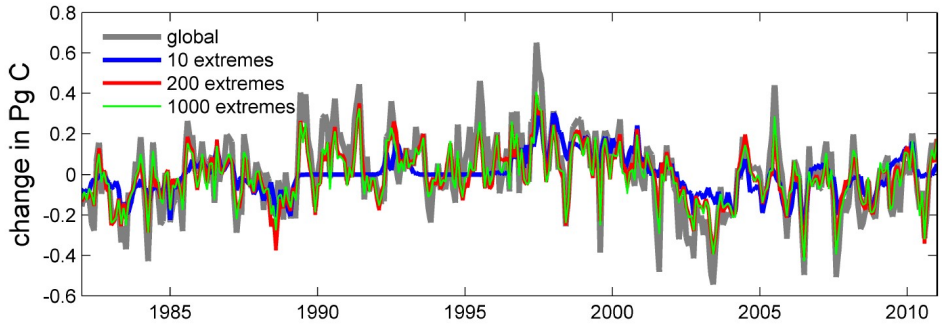

Figure 1.2: Global GPP anomaly (gray); 10, 200, and 1000 largest positive and negative 10th percentile extremes in GPP (blue, red, and green lines, respectively), on a monthly time scale. It can be seen that the biggest extremes largely determine the global anomaly of GPP in Pg C. Reproduced from [2]

Furthermore, Zscheischler et al. found that negative GPP extreme events are generally larger than positive extreme events. This asymmetry could be due to negative events such as droughts and fires having instantaneous effects, while the effects of positive events take place on a longer timescale and are therefore partially undetected.

Finally, various drivers, or possible origins, of extreme reductions in GPP were investigated. These include extreme temperatures, extreme precipitation, droughts and fires. It is found that negative GPP extremes are most often associated with anomalous low values of water availability. Extreme temperatures and extreme precipitation have a comparatively small role. Most of the negative GPP extremes could be explained with the use of these four drivers. Other causes for these GPP anomalies might lie in pest outbreaks, extreme winds and human deforestation.

1.3 Effects of extreme weather events on the carbon balance

In the previous section it was mentioned how extreme weather events (EWE) such as droughts, heatwaves, extreme precipitation, extreme low temperatures and storms, are associated with GPP extremes. In this section the effects of such EWE on the carbon cycle are further explored. The effects of EWE on the carbon cycle can either occur during or after the EWE, where its impact can be either direct or indirect. Direct impacts are directly caused by the EWE if a certain resilience threshold is passed, while indirect impacts are a cause of the increased susceptibility of the ecosystem to future EWE [15].

1.3.1 Droughts

Droughts can have various direct and indirect impacts on the carbon cycle. It is agreed upon that droughts significantly reduce terrestrial ecosystem carbon sinks, and may even turn them into carbon sources. Droughts may substantially reduce vegetation productivity in most regions as the reduced water availability limits plant growth [16]. Droughts may cause plants to make various structural or physiological adjustments that decreases their CO2 assimilation rate, such as stomatal closure and changes in leaf area. Severe and persistent droughts can not only reduce the vegetation productivity but also cause vegetation mortality. This may cause the ecosystem to regress or even collapse, resulting in lasting effects [17].

An indirect effect of droughts is the increased susceptibility to forest fires. This is caused by the reduced moisture content of the trees increasing its flammability, thus increasing the probability and spatial extent of a forest fire. Forest fires release large quantities of carbon to the atmosphere and may

have lasting effects on the ecosystem through the change in vegetation and soil structure [18].

1.3.2 Extreme high temperatures/heatwaves

Previous studies have shown that extreme high-temperature events or heatwaves often reduce the GPP of terrestrial ecosystems [19]. The effects of these events on plants range from disruptions in enzyme activity, affecting photosynthesis and respiration, to changes in growth and development. Extreme high temperatures and droughts are often connected to each other and initiate a positive feedback mechanism due to suppressed evaporative cooling caused by soil moisture deficits [15]. The timing and duration of extreme high-temperature events play an important role on its impact on the carbon cycle. Unusually warm temperatures at the end of the winter can cause a ”false spring”, which will induce plant activity too early and make them more susceptible to frost events [20]. In general, extreme temperatures events during the growing season have the biggest impact on the carbon cycle, while for events that happen in the rest of the year the impact is small [15].

The amount of insect and pathogen outbreaks are indirectly affected by extreme high temperature events in combination with droughts. Warmer temperatures are more favourable for the increase in population while soil water deficits caused by droughts make the trees more susceptible to such outbreaks [21].

1.3.3 Extreme precipitation

Extreme precipitation events can change the CO2 fluxes in the soil and CO2 uptake by plants, cause erosion in the top layers of the soil and lead to floods resulting in tree mortality. The impact of such events depend on the season and the biome type. Changes in precipitation during the growing season have a greater impact on the carbon cycle than changes during non-growing seasons. In arid regions, extreme precipitation will increase the soil water availability and therefore enhance the productivity. On the other hand, extreme precipitation will have a negative effect on productivity and carbon sinks in more humid regions [22].

1.3.4 Extreme low temperatures

Extreme low temperatures that occur during the growth season can slow vegetation development. Most of its impacts are events associated with frost. Freezing can cause damage to plant tissues and even result in their death. Extreme low temperatures may result in ice storms, where precipitation liquid freezes after coming in contact with vegetation. The added weight of the surrounding layer of ice will result in the loss of branches or even uproot entire trees [15].

1.3.5 Extreme wind

Extreme wind and tropical cyclones are often associated with extreme precipitation events and together they can cause severe damages and decrease the productivity, for example by soil erosion. Hurricanes can kill or damage a massive amount of trees which will then be converted to CO2 and returned to the atmosphere by either decomposition or fires. This is then followed by a decrease in productivity in the following years; while small events in fast-growing regions recover quickly, larger events in slow growing regions can take more than a century to recover from [23].

1.3.6 Lagged effects

It is important to realise that EWE may have impacts years after their occurrence. This will cause ecosystems to have an altered response to subsequent EWE. Drought and heat-related events can cause lasting damages to plants, especially in regions with low annual precipitation.

It is important to define a time scale of the overall effect of an EWE on the carbon cycle. This way one knows how much of a time scale needs to be taken into account when studying the effects of an EWE on the carbon cycle. Recent studies have shown that lagging effects span 1–2 years in shrubs and grasses and up to four or more years in forests [17]. Of course this is also dependant on the timing and the intensity of the event. Negative effects may in the long run be balanced by enhanced growth during recovery, depending on the resilience of the ecosystem [15].

1.3.7 Power-law behaviour in extreme weather events

It can be seen that EWE can affect the carbon cycle in various ways. Now, referring back to the power-law for GPP extremes, it would be interesting to look into some examples of power-laws associated with EWE.

It is found that areas burned by wildland fires in different places in the USA and Australia follow a power-law with an exponent between 1.3 and 1.5 [24]. However, there have also been some claims that burned areas in other regions better follow a lognormal distribution than a power-law [25]. All in all, the classification of a power-law distribution for fires is debated but it does hold up in some instances.

There were also power-laws found in various parameters associated with rainfall [26, 27]. The event size, or the amount of rain that falls during a rainfall event, is found to follow a power-law with an exponent of 1.4 for several order of magnitudes when looking at the distribution in the amount of events per year. The duration of and between rainfall events is also found to follow a power-law with an exponent of 1.6 and 1.4 respectively.

One temperature related example would be the persistence, characterized as the auto-correlation of temperature variations separated by a certain amount of days, which follows a power-law with an exponent close to 0.7 [28]. Finally, an example associated with wind would be the energy associated with tropical cyclones, which is found to follow a power-law for some part of the distribution [29].

1.4 Background on power-laws

There are various measured quantities in both natural and man-made systems that are deemed to follow power-laws. This power-law relation for some variable x can be given by:

$$p(x) = c \cdot x^{-\alpha}, \tag{1.2}$$

with some constant $c > 0$ and scaling parameter $\alpha > 0$.

A power-law distribution is heavily skewed to the right. This means that the bulk of the distribution has relatively small values while a small amount of the distribution has really high values which produce a long tail to the right of a histogram. When a histogram is plotted on logarithmic horizontal and vertical axis, it will appear to follow a straight line. Another property of a power-law distribution is that it is ”scale-free”. This means that multiplying a power-law distributed variable x by some factor b does not change the shape of the distribution, it will only change by a multiplicative constant. In figure 1.3 an example of a power-law can be seen.

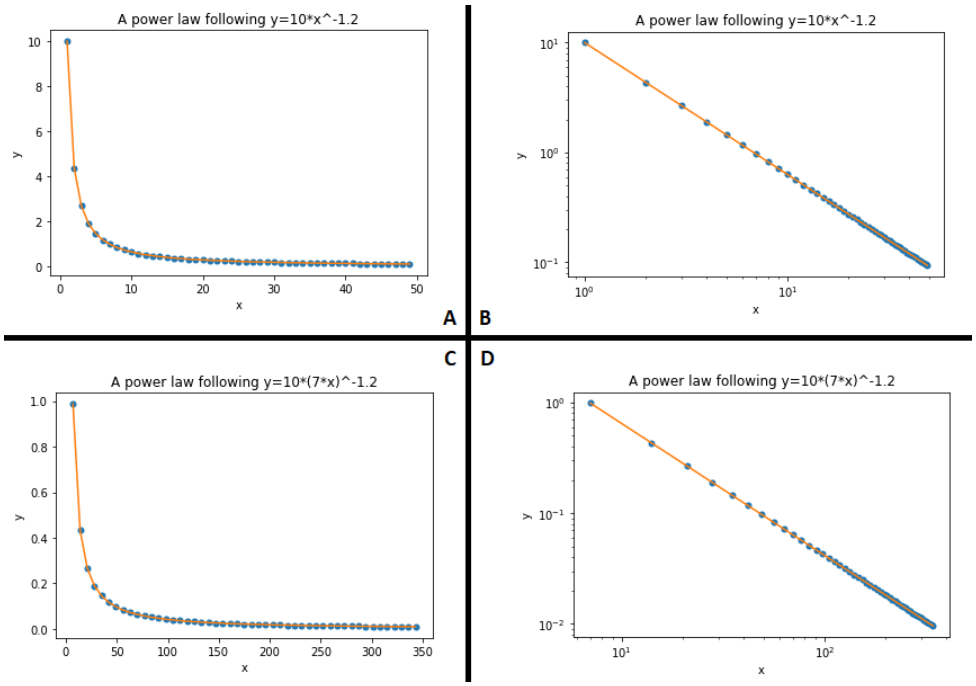

Figure 1.3: **a)** A power-law plotted with $y = c \cdot x^{-\alpha}$ with $c = 10$ and $\alpha = 1.2$. **b)** The same power-law is plotted on logarithmic horizontal and vertical axis and follows a straight line. **c)** A power-law plotted for $y = c \cdot (b * x)^{-\alpha}$ with $c = 10$, $b = 7$ and $\alpha = 1.2$. It can be seen that the only difference lies in the multiplicative constant and that the overall shape of the curve does not change. **d)** Also with logarithmic horizontal and vertical axis, the overall shape of the curve stays the same

1.4.1 Percolation theory

There are several ways in how a power-law distribution can be generated. Some of them are quite complex but there are also some simple algebraic methods for generating power-laws. One of the mechanisms for power-law generation is that of critical phenomena, where a system will follow a power-law distribution if it is in a certain ”critical” state. One example of this kind of critical behaviour can be found in percolation theory. Percolation theory has its applications in a broad range of subjects and fields such as the understanding of networks[30], earth topography[31][32] and magnetic models[33].

Percolation theory can be explained with the help of a square lattice, one where every square can either be ’occupied’ or ’empty’. Every square has a probability p to be occupied, independent of whether its neighbors are occupied or empty. If occupied squares neighbor other occupied squares, they are said to be connected and form clusters. An isolated occupied square would have a cluster size of *one*, while a cluster consisting of s clusters would have cluster size s. The size of the clusters depend on p. If p is small there will be a lot of isolated occupied squares and small clusters consisting of only a few occupied squares. On the other hand, if p becomes close to unity nearly all occupied squares are connected to each other and form a large cluster extending from one end of the lattice to the other end, as can be seen in figure 1.4. In an infinitely sized lattice, this cluster would be infinite in size and is therefore called an ’infinite cluster’. There is a certain value for p for when such an infinite cluster appears, this is called the percolation threshold p_c. This percolation threshold is not universal, it depends on the type of lattice that is considered. The percolation threshold for a simple cubic lattice for example is 0.307 [34].

At the percolation threshold the size of the clusters scale with a power-law with an exponent τ, in a similar way as the size distribution of GPP extremes that scaled with α as discussed in section 1.2. τ has

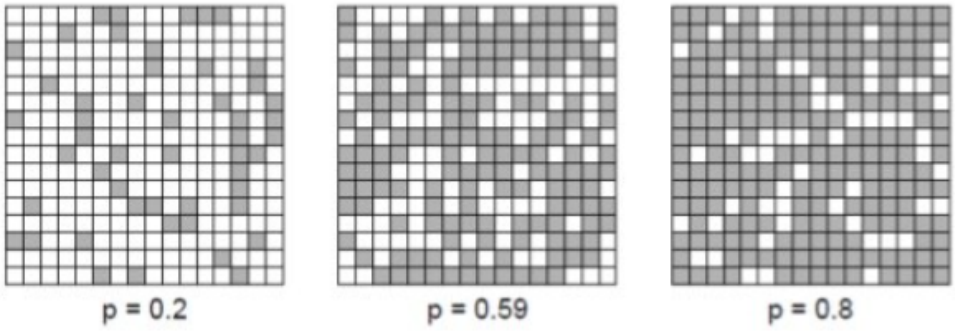

Figure 1.4: Percolation in a square 2D lattice visualised: In the left image $p < p_c$ so there are a lot of isolated occupied squares and small clusters. In the middle image $p = p_c$ so there is an 'infinite cluster' going from one end of the lattice to the other end. In the right image $p > p_c$ and all the occupied squares are part of one big cluster. Reproduced from [3]

.

a value of 2.19 for all lattices in 3D [35]. The number of clusters n_s of size s can be displayed as follows [36]:

$$n_s \sim s^{-\tau}. \tag{1.3}$$

This exponent τ is related to the earlier described exponent α as $\tau = \alpha + 1$. The size distribution, of which α is the related exponent of, describes the sum of all clusters larger than s. This is in principle the integral of n_s, thus the raise of the exponent by 1 from α to τ. Proof of this can be seen in A.1.

When $p \neq p_c$ this relation changes to:

$$n_s \sim s^{-\tau} \cdot f(s/s_\zeta), \tag{1.4}$$

with

$$f(s/s_\zeta) \sim e^{-s/s_\zeta}, \tag{1.5}$$

It can be seen that for $p \neq p_c$ there is a mix of a power-law and an exponential in the distribution of the clusters. The distribution of clusters larger than a certain size s_ζ will not behave like a power-law anymore but like an exponential. This characteristic size is related to the distance to the percolation threshold by [37]:

$$s_\zeta \sim |p - p_c|^{-1/\sigma}, \tag{1.6}$$

with σ being an exponent describing the divergence from p_c.

There are some similarities between the clusters formed in percolation theory with the clusters of GPP extreme events. In both cases clusters are formed out of one part of the data, this part being "occupied squares" for percolation theory and "extremes" for GPP. Thus, it can be hypothesised that the power-law that Zscheischler et al. found in GPP extreme events can be explained using percolation theory. The analogy between GPP extremes and percolation theory is further explored in section 4.1.

1.5 Summary and further outlook on the setup of this thesis

The power-law in the size distribution of GPP extreme events is an interesting phenomenon. In order to understand how this power-law came to be, it is first necessary to replicate the results of Zscheischler et al. Thus, in the first few sections of my thesis, I will do an analysis of extremes in the same way as Zscheischler did in his work. I will use data obtained from CMIP6 (Coupled Model Intercomparison Project Phase 6) simulations performed with the Max Planck Institute for Meteorology Earth System Model with a spatial resolution of 1.0° that spans from 1850-2350. This data will be pre-processed by subtracting the trend and seasonality. From the resulting GPP anomalies, clusters of extremes are searched out, which are based on a certain percentile of occurrence. The distribution of the GPP size of these clusters will then be plotted on a double logarithmic scale. Here a straight line is expected, indicating a power-law.

One caveat of the power-law found by Zscheischler is that it only exists for roughly two orders of magnitude. One reason for this could be the lack of data. Therefore it is interesting to see if a larger power-law region could be found with the data from the control simulation which consists of a much longer time period.

The origins of the power-law in GPP extremes could be in the GPP distribution itself. This hypothesis will be tested first. The origins of the power-law could also have a more mathematical background, based on mechanisms behind the power-law and independent of the type of data. In section 1.4.1 percolation theory came up as a hypothesis that could explain this power-law behaviour. GPP extremes will be analyzed in the framework of percolation theory, to find out if this hypothesis holds up.

In section 1.3, it could be seen that there are various ways in how EWE can affect the carbon cycle. Droughts, heavy precipitation, extreme high and low temperatures, and extreme wind all affect GPP in their own ways. Therefore, it could be assumed that some of the power-law behaviour in GPP extremes can also be seen in extremes in temperature and precipitation. This will be analysed by looking at the distribution of extreme events in temperature and precipitation and analyze if similar patterns occur to that of extreme events in GPP. Temperature and precipitation data will be handled in the same way to find the connection between GPP and EWE.

All in all, it can be concluded that there are several possible explanations on how the power-law in the distribution of GPP extremes is generated. It could have its origins in GPP itself or its origins could be mathematical and independent on the type of data and its physical meaning. In this thesis I will try to find this out with the methods described above, to ultimately be able to answer the question: ”What is the origin of the power-law behaviour in the size distribution of GPP extreme events?”

Chapter 2

Methods

2.1 Methodology

Here I will explain the data and the methods that are used in the subsequent sections to identify extreme events in GPP. First I will present the data that was used and then the methods in finding extreme events.

2.1.1 Data

The data that was used for the detection of extreme event in GPP is generated by a model named MPI-ESM1.2-HR which is part of Coupled Model Intercomparison Project Phase 6 (CMIP6) [38]. MPI-ESM is a model that couples the atmosphere, ocean and land surface through the exchange of energy, momentum, water and carbon dioxide [39]. The data spans a time period of 500 years with a spatial resolution of 1.0° in latitude and longitude. The advantage of using data of such a simulation instead of real-world data is the larger amount of data due to the much longer time period. This is important as there is a sufficient amount of data needed to analyze extremes, as extremes by definition only consist of a small part of the data. In this thesis the focus is put on the tropics, meaning the latitude area between 30°N and 30°S. In this region the GPP values will generally be the highest and therefore the largest GPP extremes will be found here, thus making it a more interesting region to look at than other latitude areas.

2.1.2 Finding extreme events

There are three steps in finding and recognizing extreme events from the GPP data. The first step is to preprocess the data to be better able to compare data across different seasons and years. Then, extreme values of GPP have to be found before identifying extreme events. These three steps follow the methods of Zscheischer et al. [1].

Preprocessing of the data

Before looking into extremes in GPP, the data needs to be preprocessed. Often with Earth observations, most datasets are expected to have some kind of seasonality and (non-)linear trend [40]. Subtracting the seasonality and the trend allows for a better comparison in values and extremes across time without the influence of variations in different seasons and years.

For the GPP data, first the trend is removed. This is done by calculating the linear trend for every pixel separately and then subtracting it from the raw GPP data. After the trend, the seasonality is removed. This is done by calculating the mean value for every pixel for every month, then subtracting it from the

Table 2.1: An overview of the used preprocessing steps for GPP.

Preprocessing step	**Method**	**Goal**
Remove trend	Subtract linear trend from each pixel	Comparability across time
Remove seasonality	Subtract monthly mean from every pixel	Comparability across seasons

trend removed GPP data. The resulting preprocessed data now describes the variation of GPP compared to the mean and will therefore be referred to as 'GPP anomalies' from now on. An overview of these preprocessing steps can be seen in table 2.1.

Extremes

Extremes can be defined as the occurrence of certain values in the tails of the probability distribution of the GPP anomalies. The definition of extremes here is based on percentiles of the values of the GPP anomalies. Extremes are values that occur less than $n\%$, n being the percentile that is used. For example: 10th percentile extremes are values that occur less often than or equal to 10% of the time. Thus for GPP extremes, 10th percentile extremes are either the lowest 10% (negative extremes) or the highest 10% (positive extremes) of the GPP anomalies. An illustratory image of 10th percentile extremes within a timestep can be seen in figure 2.1.

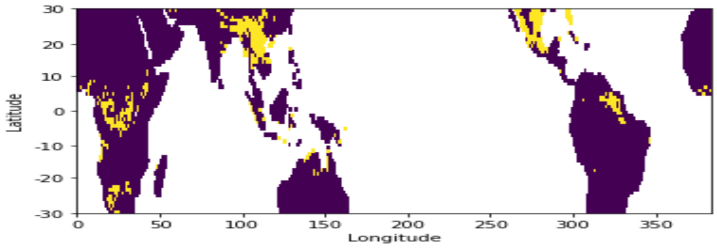

Figure 2.1: An example of the spread of 10th percentile extremes within a timestep. Extremes are marked as yellow while all land is marked as purple.

Extreme events

The first step in identifying extreme events is labeling extreme voxels, voxels that contain an extreme value for a certain percentile. The next step is finding connected components between the labeled extreme voxels. Connected voxels then form together a cluster, similar to those described in section 1.4.1. Such a cluster will be called an 'extreme event'. Whether two voxels are considered neighbours or not depends on the 'connectivity' that is used: a connectivity of 6 means that only horizontal and vertical connections are considered, a connectivity of 18 means that diagonal connections are also considered and a connectivity of 26 means that all connections surrounding the voxel in a 3x3x3 data-cube are considered. A visual representation of this can be seen in figure 2.2. Extreme voxels (cluster size $s = 1$) that are not connected to any other extreme voxels are not considered an extreme event. Finally, the size of an extreme event is then determined by the integral of the corresponding GPP anomalies in time and space

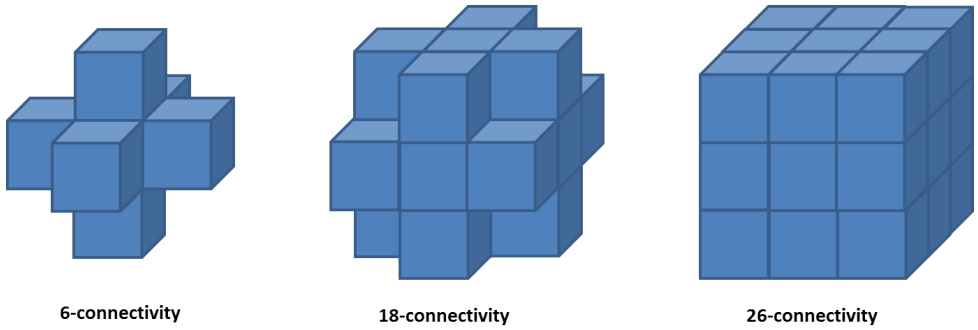

Figure 2.2: A visual representation of 6-,18- and 26-connectivity. Reproduced from [4]

.

within a cluster, in units of grams carbon. For negative extremes the absolute value is taken, in order to be able to better compare negative and positive extremes.

2.2 The size distribution of GPP extremes

In figure 2.3a the size distribution of GPP extreme events can be seen for 5th percentile extremes with a connectivity of 26, plotted on double logarithmic axis. As there are both a similar amount of negative (neg) and positive (pos) extremes, they fit well into the same plot without having to be normalized. At the top of the figure the total amount of negative extremes is given. The figure contains a cumulative size distribution where the x-axis gives the size of an event x and the y-axis gives the number of events that are larger than x. This is the same type of plot as the one from Zscheischler et al. from section 1.2, which is shown again in figure 2.3b. The difference between the two plots being that Zscheischler et al. used a normalized y-axis describing the fraction of event larger than x, while in my plot the cumulated number of events is shown on the y-axis.

Both curves start with a somewhat linear regime. This linear regime is where the power-law occurs and will therefore be referred to as 'power-law region' from now on. After the power-law region there is a drop off at the largest few events. This drop-off presumably occurs because of the size largest extreme events being limited by the continental borders. The fact that both figures have similar characteristics demonstrates that the data from the CMIP6 simulation behaves similarly to the observational data from Zscheischler et al. and is a justification for using such simulated data for this study.

The most notable difference between the two figures is the difference in the width of the power-law region. While in the plot of Zscheischler et al. this power-law region covers about 1 order of magnitude, in my plot it can be seen that for both negative and positive extremes there is a power-law region that covers more than 2 orders of magnitude. This difference may be caused by the difference in the amount of data: whereas the observational data from Zschseischler et al. spans a 30 year period, the simulation data from CMIP6 spans a period of 500 years. This larger power-law region and the larger amount of data means that the scaling behaviour in this power-law can be more reliably studied.

2.2.1 The size distribution for 10th and 1st percentile extremes

Next, there will be taken a look at size distribution of GPP for different percentiles. More specifically, the 10th and 1st percentile extremes will be looked at, as those are the other percentiles that were studied by Zscheischler et al.

(a) (b)

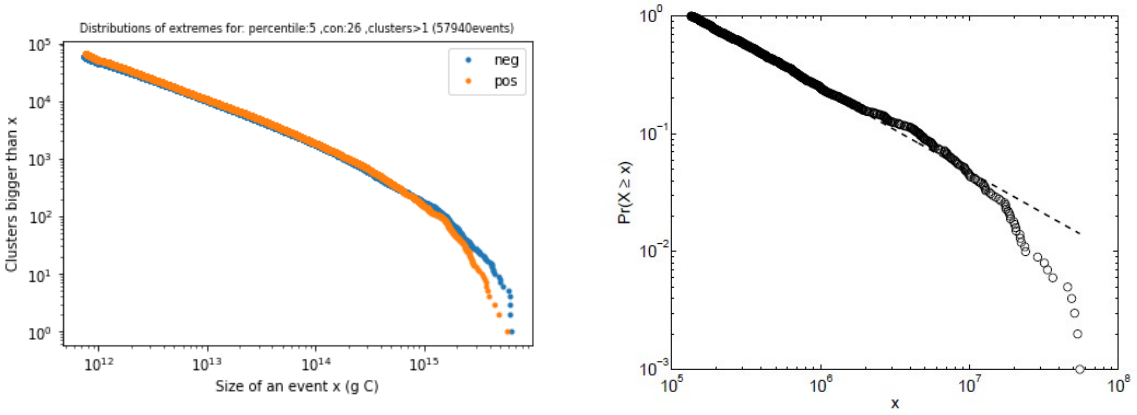

Figure 2.3: Size distribution of GPP extremes for the 5th percentile and 26th connectivity **a)** Produced using CMIP6 data. 'neg' stands for negative extremes, 'pos' stands for postitive extremes. **b)** From Zscheischler et al. Reproduced from [5]

The size distribution of GPP extreme events for 10th percentile extremes with a connectivity of 26 is plotted on double logarithmic axis in figure 2.4. There is a power-law region present for event sizes of almost 4 orders of magnitude, with a scaling parameter of $\alpha = 0.75$ and $\alpha = 0.73$ for negative and positive extremes respectively.

In Figure 2.4b, again the same size distribution of GPP extreme events can be seen plotted on a logarithmic y-axis. There is also a linear region present here, seemingly in the region where the drop-off occurred in the double logarithmic plots. This linear region indicates an exponential distribution. The linear region here however covers much less events and also only covers around 1 order of magnitude in terms of event sizes.

(a) (b)

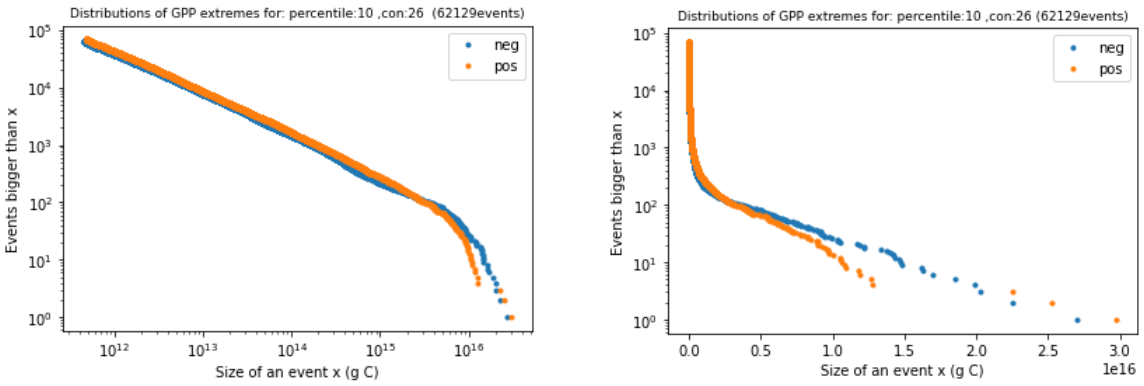

Figure 2.4: Size distribution of GPP extremes for the 10th percentile and 26-connectivity for **a)** logarithmic x- and y-axis, **b)** logarithmic y-axis

When looking at 1st percentile extremes in figure 2.5a, it can be seen that on first look the broad characteristics of the curve are the same to the 5th and 10th percentile extremes, but on closer look there are clearly some differences. It is not clear whether there is something that could be called a power-law region. In general it can be seen that there is less variety in the event sizes for the 1st percentile case covering about 3 orders of magnitude in total while the event sizes differ in about 5 orders of magnitude

(a) (b)

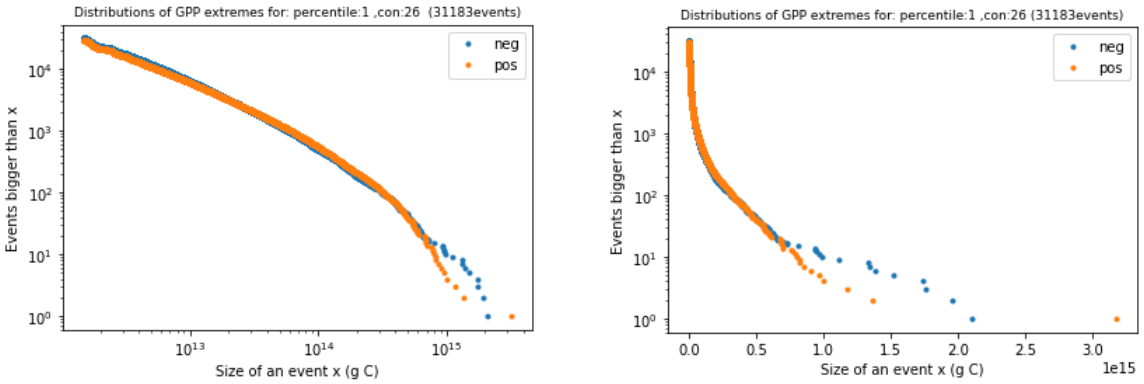

Figure 2.5: Size distribution of GPP extremes for the 1st percentile and 26-connectivity for **a)** logarithmic x- and y-axis, **b)** logarithmic y-axis

for the 10th percentile case. The scaling parameter also has different values with $\alpha = 1.10$ and $\alpha = 1.03$ for negative and positive extremes respectively. Looking at the single logarithmic plot in figure 2.5b, there is also a linear region present although it is also smaller and less clear than in the 10th percentile case because of the scarcity of data in the tail of the distribution. One could say that the linear region covers about half an order of magnitude before the data gets too scarce to take any conclusions of.

For both the 1st and 10th percentile GPP size distribution, the negative and positive extremes have a similar size distribution and scaling parameter. The largest events in the size distribution of negative extremes are however slightly larger than for positive extremes. For 1st percentile extremes the existence of a power-law region was not clear, while for 10th percentile extremes there was a clear power-law region for several orders of magnitude. Therefore 10th percentile extremes will be used in the showcase of size distributions from now on.

Chapter 3

Does the origin of the power-law lie in the GPP distribution?

It could be seen that there is a power-law occuring in the size distribution of GPP extremes. The question is now where this power-law comes from. The first step towards answering this question may lie in the GPP distribution itself. The occurrence of a power-law in the distribution of GPP could be an explanation of the occurrence of a power-law in its clusters. The following hypothesis can be made: the power-law in the size distribution of GPP extremes can be explained by similar power-law behaviour in the size distribution of GPP itself. To test this, the size distribution of GPP will be looked at, before and after preprocessing, to see if there is a power-law region to be detected.

In figure 3.1a the size distribution of GPP is plotted. It can be seen that there is a large almost horizontal region that quite suddenly transitions in a near vertical drop. This indicates that apart from a few lower values, most GPP values are quite close to each other and around 10^{12}. A look at a histogram in figure 3.1b confirms that indeed most values are concentrated around 10^{12}. The horizontal region caused by a small number of lower GPP values can be explained by the existence of desert regions, most notably the Sahara, which have much lower GPP than other regions in the tropics.

When looking at the size distribution of GPP anomalies in figure 3.2a, there is similar behaviour to be seen as with the original GPP distribution. There is a horizontal region followed by a near vertical drop, although the transition between those two areas seem to follow more smoothly. Figure 3.2b shows a ”zoomed-in” version of the right side of the previous plot where the shape of this vertical region can be

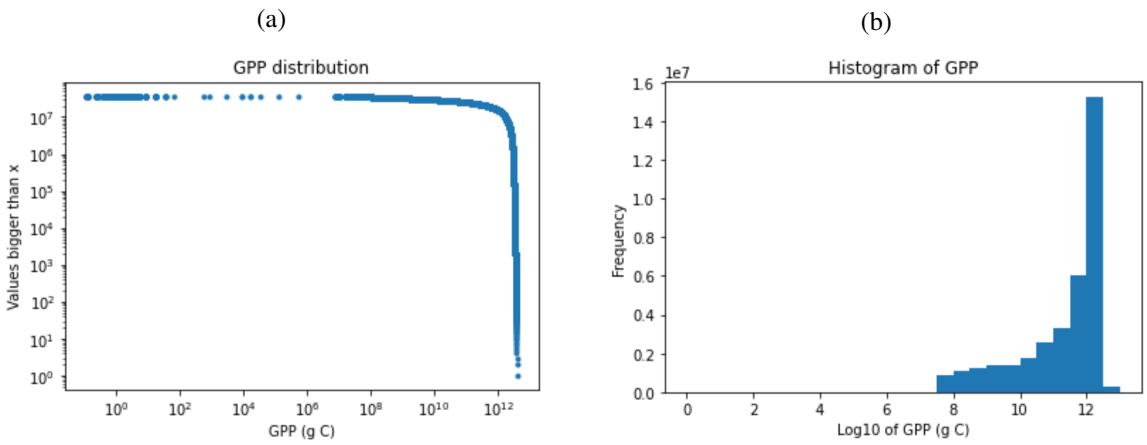

Figure 3.1: **a)** The size distribution of GPP **b)** A histogram of the size distribution of GPP

(a) (b)

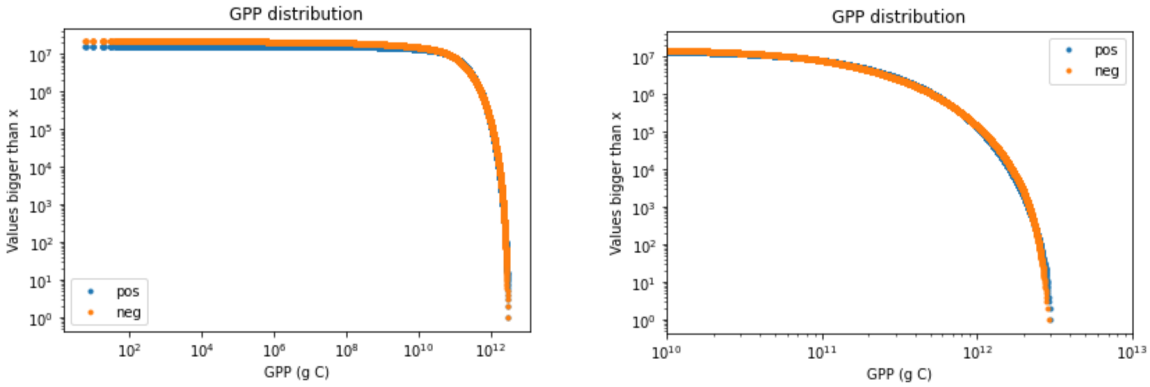

Figure 3.2: **a)** The size distribution of GPP anomalies **b)** A zoomed in version of the same size distribution

better seen. It is evident that there is no linear region and thus there is no power-law to be found in the size distribution of GPP. Thus, the origin of the power-law does not lie in the distribution of GPP itself.

Chapter 4

Can percolation theory explain the power-law scaling?

4.1 Percolation and GPP extremes

It is clear that the origin of the power-law is not to be found in the GPP distribution itself. This means that there should be another explanation for this power-law, unrelated to the characteristics of GPP itself. An analogy with percolation theory might just be able to provide this explanation.

The clusters of GPP extreme events share some similar characteristics with clusters formed in percolation theory. In both cases clusters are formed out of part of the data, this part being 'extremes' for GPP and 'occupied squares' in percolation theory. This part of the data is defined by a percentile for GPP which is analogous to the probability p that is used in percolation theory, with for example the 1st percentile corresponding to $p = 0.01$. Also in both cases a power-law is present that is related to the size of the clusters. The exponents for this power-law, α for GPP and τ for percolation theory, are related to eachother by the relation $\tau = \alpha + 1$ as can be seen in section A.1. Therefore the subsequent investigations starts out from the hypothesis that the power-law in the size distribution of clusters of GPP extremes has a link to percolation theory.

4.1.1 Percolation threshold

One important parameter for percolation theory is the percolation threshold p_c. p_c for different values according to percolation theory are obtained from Shklovskii et al. [7]. These values can be seen in table 4.1.

The main characteristic of p_c is the occurrence of an infinite cluster. Since the grid that is used is not infinite, an infinite cluster here is defined by a cluster that encompasses all 6000 time steps. The lowest percentile where such an infinite cluster can be found is then defined as p_c. The following values for p_c can be found for the different connectivities: $p_{c6} = 26\%$, $p_{c18} = 20\%$, $p_{c26} = 20\%$.

Table 4.1: Percolation threshold according to percolation theory. Values obtained from Shklovskii et al. [7]

Con.	6	18	26
p_c	31%	14%	10%

As seen earlier in section 1.4.1, finding α at this point and adding 1 to them gives values for τ: $\tau_{c6} = 1.66$, $\tau_{c18} = 1.72$, $\tau_{c26} = 1.71$.

It can be seen that there is some difference in the τ values between the connectivities. Moreover, all of these values deviate a bit from the value obtained from percolation theory of 2.19.

4.1.2 Difference between GPP extreme events and percolation theory

The values that are found for p_c in GPP extremes deviate from the numbers found in percolation theory. 6-connectivity would correspond to a simple cubic lattice, which has $p_c = 31\%$. 18-connectivity would give a value of $p_c = 14\%$ and 26 connectivity would give a value of $p_c = 10\%$.

Also in the values of the exponent τ there is some deviation from percolation theory. All three values of τ are lower than the 2.19 which is the value expected for 3D-percolation from the theory [35].

Not all assumptions underlying percolation theory apply to the clusters of GPP extreme events. There are some differences between the clustering mechanisms of GPP extremes and of percolation theory which causes the differences in them for p_c and in τ. The main differences between the two situations that are relevant can be summed up in the following three points:

1. In GPP extremes the value of a voxel is the GPP-value at a certain point in space and time. This GPP-value varies for different points in space and time and thus every voxel within GPP extremes has a different value. In percolation theory one only deals with cluster sizes, which could be viewed in the same way as if each voxel had the same GPP-value of 1.

2. There is an inter-dependence of neighbouring voxels in the GPP data while in percolation theory all voxels are independent of each other. If a certain voxel has an extreme value in the GPP data, its neighbours in time and space will have a higher than average probability of having an extreme value as well. This is due to extreme weather events such as droughts and heatwaves, which cause extremes in GPP, often being spread over an area worth several voxels in space and time.

3. The spread of the GPP data is limited to the land-areas on earth. As there are no GPP values in the ocean, there are also no voxels with extreme GPP values in the ocean. This means that the clusters of GPP extremes are limited by continental borders which makes it harder to form large clusters. In percolation theory every voxel can be occupied and there are no spaces that are empty by default.

Getting rid of these three differences may result in power-law scaling behaviour that is closer to percolation theory. Starting from section 4.3 I will step for step eliminate these differences by altering the GPP data, to test this hypothesis.

4.2 Scaling parameter

The scaling parameter α is an important property of the power-law. The scaling parameter is calculated by performing a linear fit on the power-law region in the double logarithmic plots for the size distribution of GPP.

4.2.1 Systematic determination of the power-law region and of the scaling parameter α

In order to study the power-law in GPP extremes, first it has to be determined in what part of the size distribution this power-law exists. It is hard to determine what exactly comprises the power-law region. The choice of its starting and ending points will always be subjective to a certain degree. It is important to be

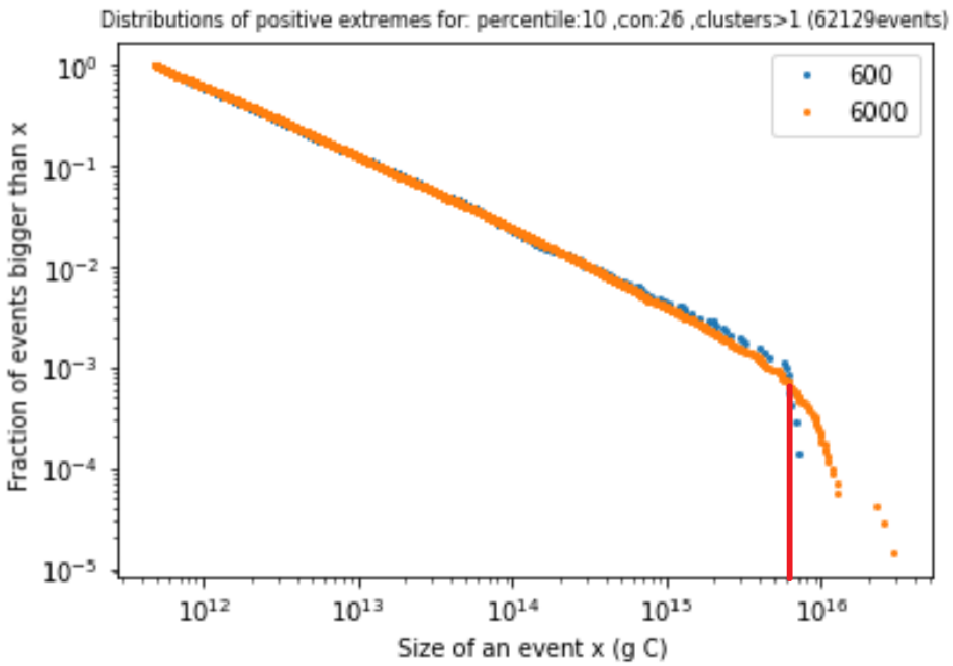

Figure 4.1: Example demonstrating how the end of the power-law regin is found. GPP6000 and GPP600 plotted in the same graph. As indicated by the red line, at an event size around $6 \cdot 10^{15}$ the two data curves start deviating from each other. This point is taken as the end of the power-law region. GPP6000 and GPP600 are both normalized to fit in the same plot.

consistent though and to choose the power-law region in a systematic way, in order to be able to compare results between different percentiles and connectivities. Here I present my method for determining the power-law region.

The starting point of the power-law region is determined by using a python function, named piecewise linear functions (see section A.2). This function finds linear regions in a graph and can therefore determine when the size distribution of GPP enters the power-law regime in the double logarithmic plot. This function does not work as well for finding the ending point of the power-law regime because of the scarcity of the data towards the end of the curve.

The determination of the ending point of the region is done by comparing GPP data from different lengths in time. The data is normalized by dividing the y-axis by the total number of events, in order to compare data from different lengths in time. The size distribution of GPP for a time periods of 600 months (GPP600) is compared to the size distribution of GPP with the usual time period of 6000 months (GPP6000). This approach has the following rationale: As the power-law region somewhat scales with the used time period, the power-law region for GPP600 is smaller than that of GPP6000. Therefore it can be assured that the region in which a power-law occurs for GPP600, GPP6000 also has power-law behaviour. Thus, the point where GPP600 starts deviating from GPP6000, indicated in figure 4.1, serves as a lower bound estimate for the end of the power-law region. The x-value of the ending point of the power-law region is determined as the point where the difference in the y-value between GPP600 and GPP6000 is larger than 0.1%.

Now with both the starting and ending point of the power-law region determined, a least squares fit is performed over this region to determine the value of α. For this, a python module is used called ”numpy.polyfit”, which is further explained in section A.3.

4.2.2 Values of the scaling parameter

As can be seen in figure 4.2, the scaling parameter α depends on both the percentile and connectivity that was used. For the lower percentiles, a relation is present where α decreases with increasing percentiles. Then, α converges to around 0.8 for 6-connectivity and 0.7 for 18- and 26-connectivity, with a slight drop at higher percentiles. Connectivity plays a smaller role than the chosen percentiles, especially the difference between 18 and 26 connectivity seems to be small.

(a) (b)

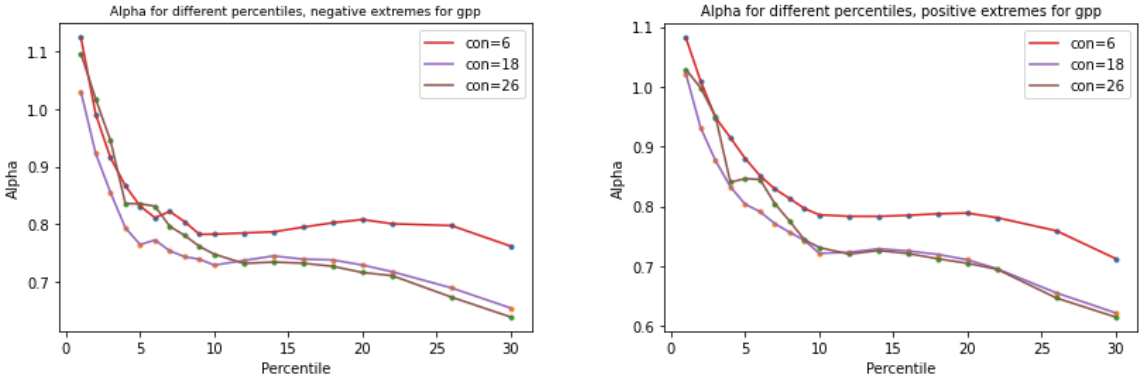

Figure 4.2: The scaling parameter α for 1st until 30th percentile and 6-, 18- and 26- connectivity, **a)** negative and **b)** positive extremes for GPP extreme events

A better way to study the behaviour of the scaling parameter in relation to percolation theory, might be to look around values of the percolation threshold p_c where critical behaviour occurs. Percentiles close to p_c provide the most reliable values of α as they have a more distinct power-law regime. As explained in chapter 1.4.1, percentiles far from the percolation threshold behave less like a power-law but more like an exponential.

In figure 4.3 values of α are plotted from $p_c - 10\%$ until $p_c + 10\%$. The values of p_c are an estimation as described in section 4.1.1. It can immediately be seen that both the percentiles and connectivities have an influence on the value of α that is obtained. 6-connectivity has higher values for α compared to 18- and 26-connectivity for which values of α are much closer to each other. The overall pattern is the same for all curves: a decline in α with increasing percentiles, especially after the percolation threshold is passed. The negative extremes have slightly higher values of α compared to the positive extremes.

4.3 Cluster size distribution

One of the caveats in the analogy between GPP clusters and percolation clusters, was the fact that the GPP value within a voxel varies while for percolation clusters each voxel has the same value, as was discussed in the first point in section 4.1.2. This problem can be solved by just considering the number of voxels that an extreme event consists of. This should produce power-law scaling behaviour closer to percolation theory, particularly in the values of τ. In other words this will be the distribution of GPP extreme events with all GPP values taken equal as "1". This can be seen in figure 4.4a. At the start of the curve, some "gaps" can be seen as a result of the more discrete nature of the cluster size in terms of voxels. This is followed by a power-law region for roughly 2 orders of magnitude with a scaling parameter of $\alpha = 0.77$ for both positive and negative extremes. This is followed by a drop-off which somewhat coincides with the linear region in the single logarithmic plot in figure 4.4b.

(a) (b)

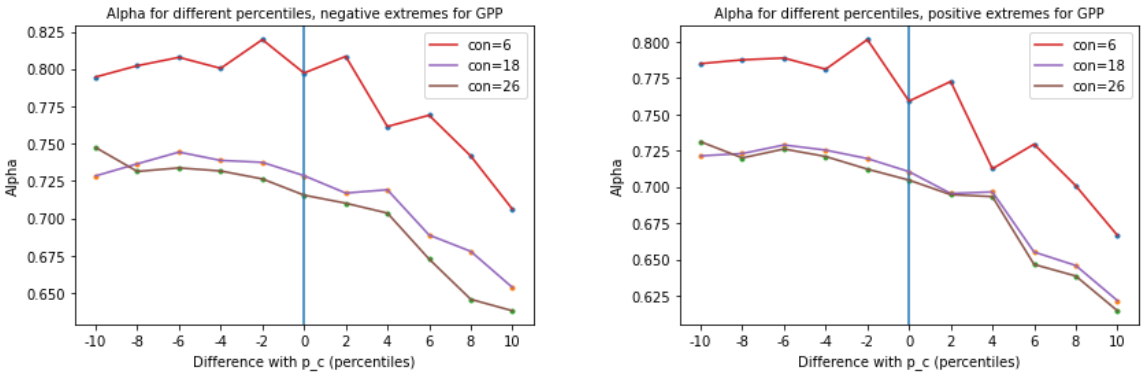

Figure 4.3: The scaling parameter α for GPP extremes, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 26\%$ for 6-connectivity, $p_c = 20\%$ for 18-connectivity and $p_c = 20\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

(a) (b)

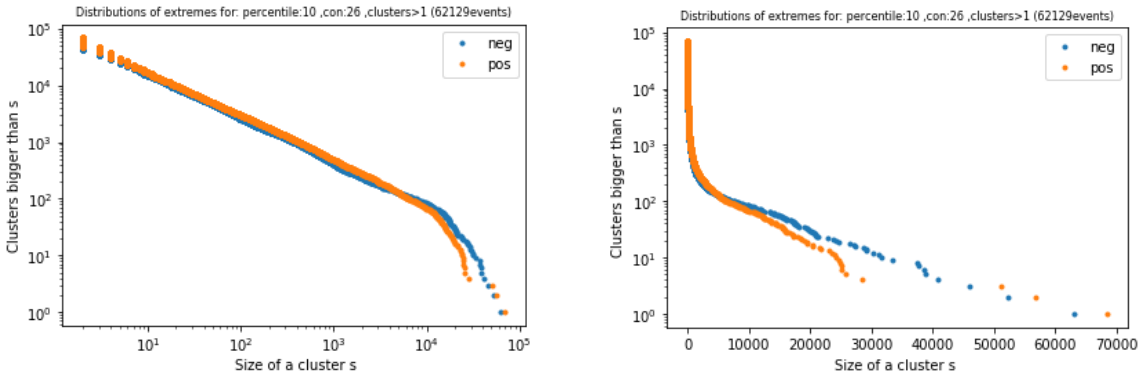

Figure 4.4: Size distribution of cluster sizes of GPP extremes for the 10th percentile and 26-connectivity for **a)** logarithmic x- and y-axis, **b)** logarithmic y-axis

4.3.1 Percolation threshold and τ

Taking the values of α at the percolation thresholds, $p_{c6} = 26\%$, $p_{c18} = 20\%$ and $p_{c26} = 20\%$, the value of the exponent τ can be estimated: $\tau_{c6} = 1.96$, $\tau_{c18} = 1.84$, $\tau_{c26} = 1.83$.

These values of τ are closer to the theoretical percolation value of 2.19. This indicates that eliminating the values of GPP, made the clusters have characteristics closer to the clusters according to percolation theory.

4.3.2 Scaling parameter

In figure 4.5, α is plotted for the percentile range of $p_c - 10\%$ to $p_c + 10\%$. There is a rise in α to be seen with increasing percentiles, especially before the percolation threshold. 6-connectivity has higher values of α than 18- and 26-connectivity. The negative extremes have slightly higher values of α than the positive extremes. Overall the values of α are higher compared to those for GPP.

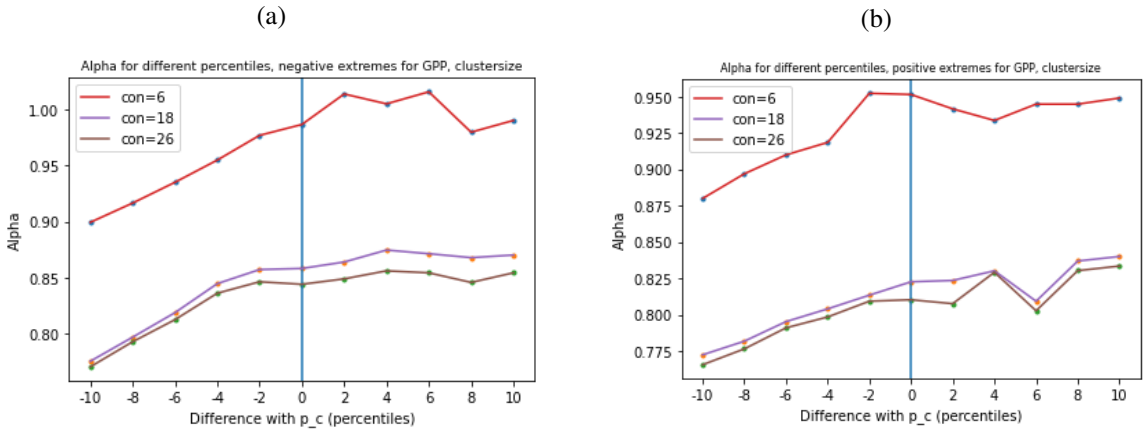

Figure 4.5: The scaling parameter α for cluster sizes of GPP extremes, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 26\%$ for 6-connectivity, $p_c = 20\%$ for 18-connectivity and $p_c = 20\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

4.4 Shuffling of data

The second difference between GPP extremes and percolation theory discussed in section 4.1.2, is that the voxels in the GPP data are not independent of each other. In this section this difference between the GPP data and percolation theory is eliminated by randomizing the GPP data and therefore destroying its correlations in time and space. This should create scaling behaviour that is closer to that of percolation theory, as percolation theory assumes uncorrelated data.

The randomizing of the data will be done by a process that is named 'shuffling'. This shuffling is a process where parts of the data are swapped around. In the following, several shuffling methods are used to randomize the data: shuffling in time, shuffling in space, shuffling in space and time, and complete shuffling. The expectation is that the more rigorous shuffling methods should give values of p_c and τ that are close to percolation theory.

4.4.1 Shuffling in time

The aim of shuffling the GPP data in time is to destroy homogeneity in time and to test the relation between GPP values in subsequent timesteps. The shuffling is done by taking a 2D slice of GPP values at a certain timestep and switching them with a 2D slice of GPP values at another timestep. The position of every voxel in space remains the same, thus the homogeneity in space remains. A good analogy for this would be the shuffling of a deck of playing cards, where every time slice of GPP would be a playing card. The shuffling is done according to the Fisher-Yates shuffling algorithm (see section A.4) [41]. A visual representation of this can be seen in figure 4.6.

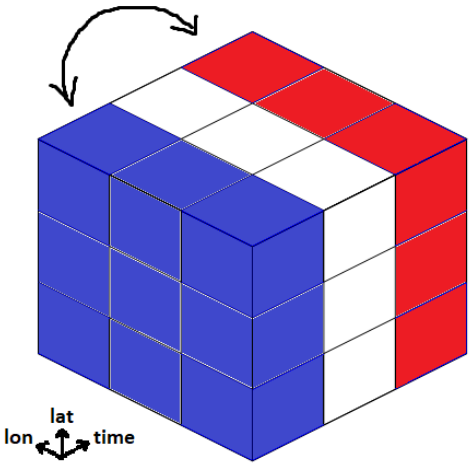

Figure 4.6: A visual representation of shuffling GPP in time. The voxels in blue are switched with the voxels in red.

Size distribution

Looking at the size distribution of GPP in figure 4.7a, a power-law region can be seen that is followed by a gradual drop-off. The power-law region covers around 3 orders of magnitude and gives a scaling parameter of $\alpha = 0.78$ and $\alpha = 0.75$ for the negative and positive extremes. The shuffling has reduced the size of the largest clusters and instead has created a lot more smaller clusters. This can be seen in the reduction of the size of the largest events and in the increase in total clusters compared to the unshuffled

Table 4.2: Values of the percolation threshold p_c and corresponding exponents τ_{gpp} for the size distribution in time shuffled GPP, and τ_{cs} for the distribution in cluster sizes. ("Con." stands for "connenctivity".)

Con.	6	18	26
p_c	27%	21%	19%
τ_{gpp}	1.87	1.81	1.82
τ_{cs}	1.98	1.89	1.88

case. Negative and positive events behave similarly although the largest negative events are slightly larger than the largest positive events.

In figure 4.7b the size distribution of the events in terms of cluster size is plotted. Similarly to the unshuffled case, it starts off with some "gaps", followed by a power-law region of 2 orders of magnitude with a scaling parameter of $\alpha = 0.81$ and $\alpha = 0.79$ for the negative and positive extremes.

(a) (b)

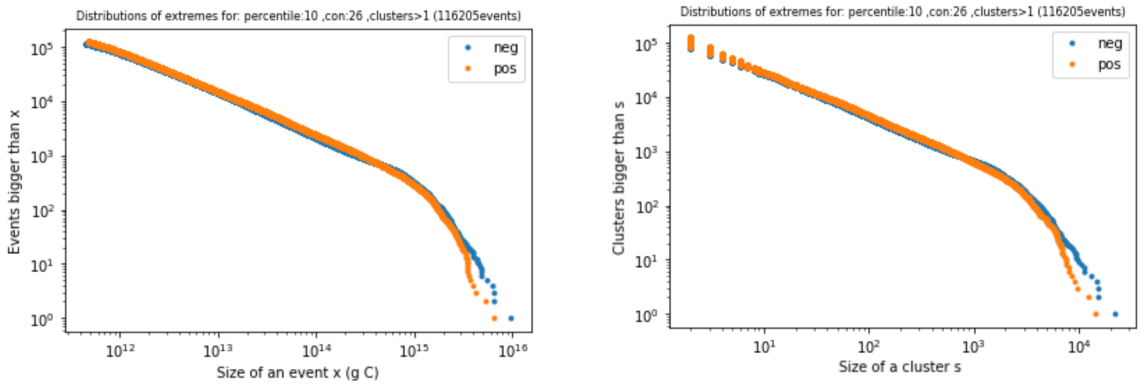

Figure 4.7: **a)** Size distribution of time shuffled GPP extremes and **b)**cluster sizes for the 10th percentile and 26-connectivity

Percolation threshold and τ

The percolation threshold p_c and the exponent τ are determined in the same way as earlier described in section 4.1.1. The results of this can be seen in table 4.2

All three values of p_c are different but still close to that of GPP before shuffling that were described in section 4.1.1. The value of p_c for 6-connectivity is higher at $p_c = 27\%$ and closer to the value of $p_c = 31\%$ from percolation theory. All values of τ are higher compared to unshuffled GPP (see sections 4.1.1 and 4.3.1), indicating that the power-law scaling behaviour in the shuffled data is closer to the that of percolation theory. It is also to be noted that τ for cluster sizes is higher than τ for GPP. Still all values of τ are quite a bit off from the usual value of $\tau = 2.19$ for 3D-percolation.

Scaling parameter

As can be seen in figure 4.8 there are some differences in the curves of the different connectivities. For 6-connectivity there is a rise in α up until $p_c + 8\%$ where there is a sudden drop. Then, for 18-connectivity there is a rise in α which transitions into somewhat of an upward arc after the percolation threshold is passed. Finally, for 26-connectivity there is an downward arc before the percolation threshold after which α rises with increasing percentile.

(a) (b)

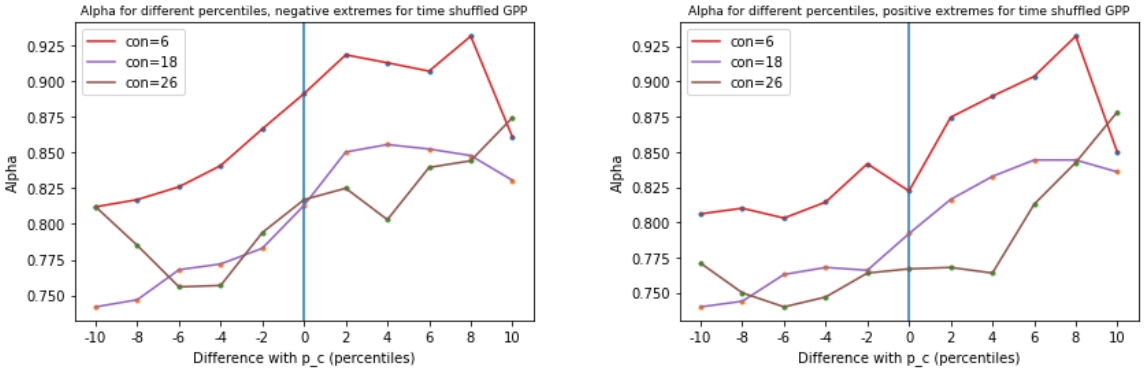

Figure 4.8: The scaling parameter α for time shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 27\%$ for 6-connectivity, $p_c = 21\%$ for 18-connectivity and $p_c = 19\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

In the case of the cluster sizes in figure 4.9, all connectivities behave similarly with a rise in α for increasing percentiles. The notable exceptions being the drop at the end of the curve for 18-connectivity and the slight decline in the beginning for 26-connectivity. Overall, the values of α for both GPP event sizes and cluster sizes are slightly higher than in the unshuffled case.

(a) (b)

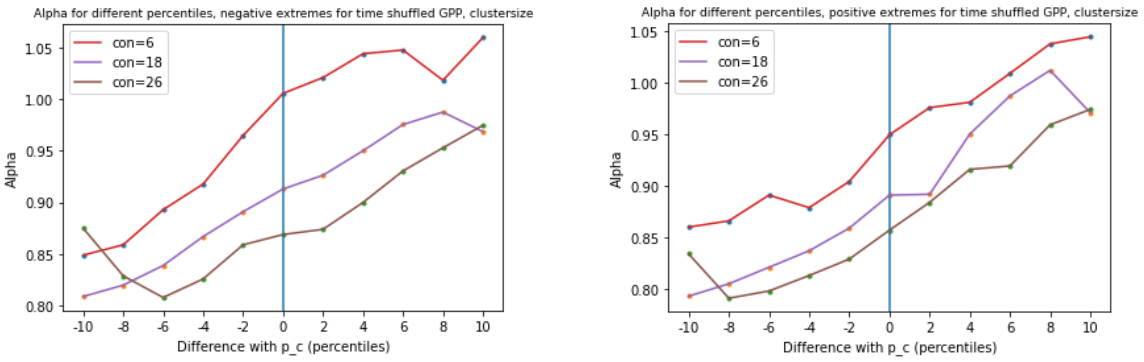

Figure 4.9: The scaling parameter α for cluster sizes of time shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 27\%$ for 6-connectivity, $p_c = 21\%$ for 18-connectivity and $p_c = 19\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

4.4.2 Shuffling in space

With shuffling in space, GPP values at different locations in space are interchanged. However, contrary to shuffling in time, all voxels do not change their location in time. A visual representation can be seen

in figure 4.10.

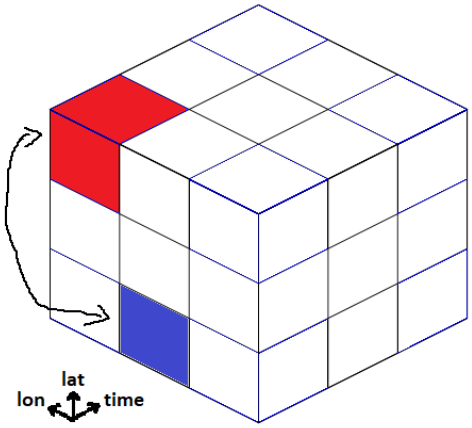

Figure 4.10: A visual representation of shuffling GPP in space. The red voxel is switched with the blue voxel that is at a different location in space but within the same time slice.

Size distribution

In figure 4.11a the event size distribution of the space shuffled GPP can be seen. There is a bit of a bending at the beginning of the curve before the transition to the power-law region. The power-law region exists for a bit more than 2 orders of magnitude and gives a scaling parameter of $\alpha = 1.04$ and $\alpha = 1.06$ for the negative and positive extremes. The largest events are bigger compared to the time shuffled GPP but smaller than those of the original GPP distribution. There are also more total clusters present compared to those other distributions, indicating a higher number of small clusters.

The size distribution of the cluster sizes is plotted in figure 4.11b. It starts again with a region containing ”gaps” followed by a linear region of 2 orders of magnitude with a scaling parameter of $\alpha = 1.10$ and $\alpha = 1.13$ for negative and positive extremes.

(a) (b)

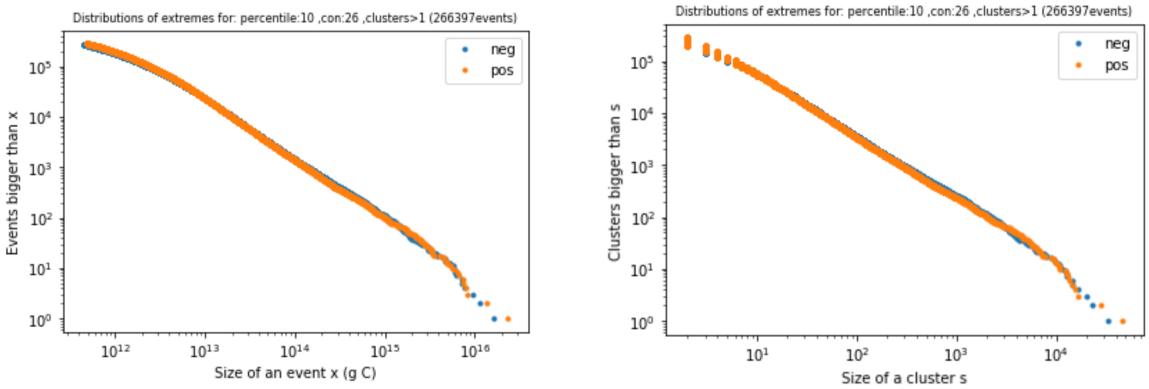

Figure 4.11: **a)** Size distribution of space shuffled GPP extremes and **b)**cluster sizes for the 10th percentile and 26-connectivity

Percolation threshold and τ

The values of p_c are lower than they were for time shuffled GPP, as can be seen by comparing the valus in table 4.3 with those from table 4.2. The difference in p_c between the connectivities is also larger than before. All values of τ are higher than the time-shuffled case. Values of τ for cluster sizes are slightly above the value $\tau = 2.19$ from percolation theory while τ for GPP are slightly under this.

Table 4.3: Values of the percolation threshold p_c and corresponding exponents τ_{gpp} for the size distribution in space shuffled GPP, and τ_{cs} for the distribution in cluster sizes.

Con.	6	18	26
p_c	25%	16%	14%
τ_{gpp}	2.16	2.04	2.06
τ_{cs}	2.22	2.28	2.28

Scaling parameter

The plots for α for space shuffled GPP in figure 4.12 have quite a different pattern than the ones that were previously seen. There is a downward arc which transitions into somewhat of an upward arc after the percolation threshold.

(a) (b)

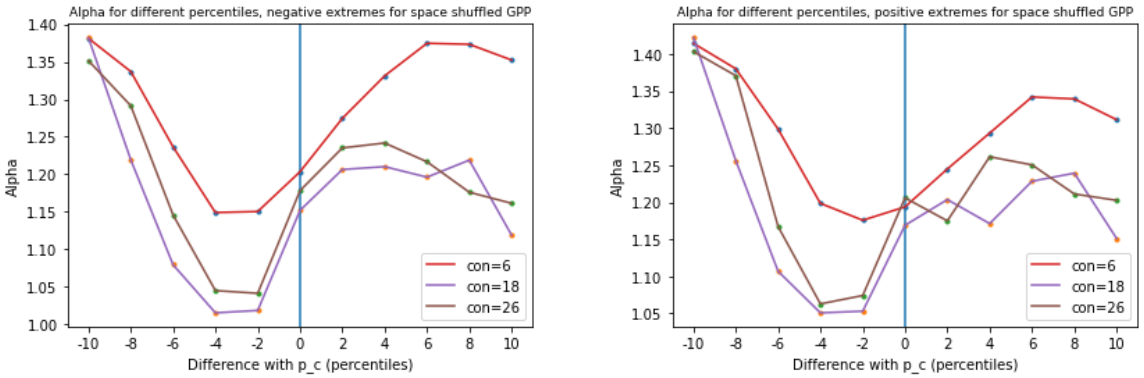

Figure 4.12: The scaling parameter α for space shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 25\%$ for 6-connectivity, $p_c = 16\%$ for 18-connectivity and $p_c = 14\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

The α for the cluster sizes in figure 4.13 follow a similar pattern. There is a difference for 6-connectivity, where there is not upward arc and instead α keeps rising after the percolation threshold. The values of α for both GPP and cluster size are much higher compared to the unshuffled case.

4.4.3 Shuffling in space and time

One has seen the result of shuffling in time and of shuffling in space. The next step would be to combine these two methods to destroy relations in both time and space, this will be called space+time shuffling.

(a) (b)

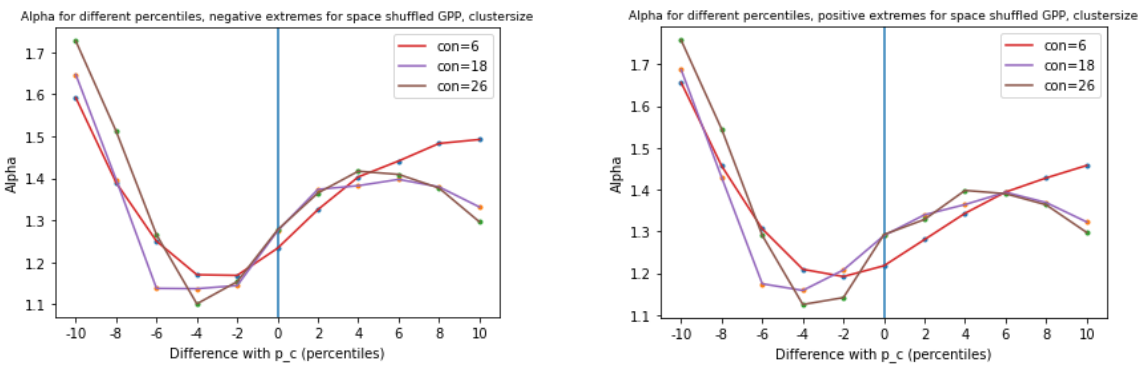

Figure 4.13: The scaling parameter α for cluster sizes of space shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 25\%$ for 6-connectivity, $p_c = 16\%$ for 18-connectivity and $p_c = 14\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

This is done by first performing shuffling in space followed by shuffling in time, using the same shuffling methods as described before.

Size distribution

The event size distribution in figure 4.14a starts with a small bending similarly to the space shuffled GPP. This is then followed by power-law region for over 2 orders of magnitude with a scaling parameter of $\alpha = 1.16$ and $\alpha = 1.17$ for negative and positive extremes. There is a slight increase in the total amount of clusters compared to the space shuffled GPP. The size of the largest events is roughly the same as for the time shuffled GPP.

The shape of the size distribution of the cluster sizes in figure 4.14 is quite similar except for the gaps in the beginning. The power-law region is slightly smaller but still covers 2 orders of magnitude with a scaling parameter of $\alpha = 1.15$ and $\alpha = 1.17$ for negative and positive extremes.

(a) (b)

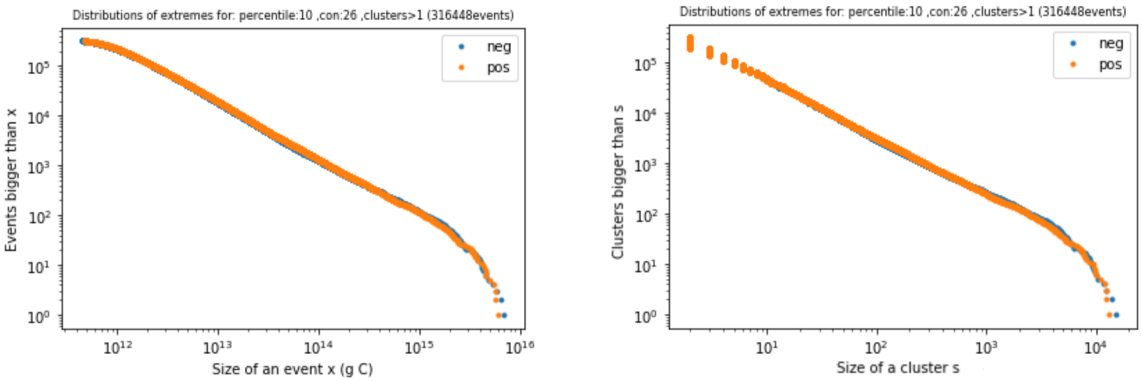

Figure 4.14: **a)** Size distribution of space+time shuffled GPP extremes and **b)**cluster sizes for the 10th percentile and 26-connectivity

Table 4.4: Values of the percolation threshold p_c and corresponding exponents τ_{gpp} for the size distribution in space+time shuffled GPP, and τ_{cs} for the distribution in cluster sizes.

Con.	6	18	26
p_c	33%	16%	12%
τ_{gpp}	2.29	2.31	2.31
τ_{cs}	2.31	2.32	2.30

Percolation threshold and τ

As can be seen in table 4.4, the differences in p_c between the connectivities is much larger than in the other cases. For 6-connectivity $p_c = 33\%$, which is higher than the theoretical value of $p_c = 31\%$. All values of τ are closer to each other and around $\tau = 2.30$. This is slightly higher than the value of $\tau = 2.19$ from percolation theory. There is not much of a difference in τ between the values of GPP and of the cluster sizes.

Scaling parameter

As can be seen in figure 4.15 for all connectivities α decreases until the $p_c + 2\%$. This decrease is sharper for the higher connectivities. After $p_c + 2\%$, α starts increasing slightly, until it converges to around $\alpha = 1.6$ for 6-connectivity and decreases slightly for 18- and 26-connectivity.

(a) (b)

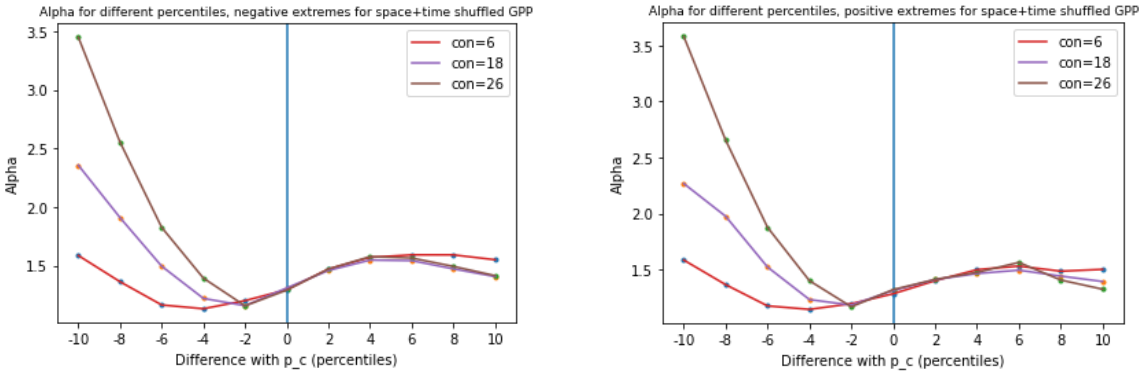

Figure 4.15: The scaling parameter α for the size distribution of space+time shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 33\%$ for 6-connectivity, $p_c = 16\%$ for 18-connectivity and $p_c = 12\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

The α for the cluster sizes in figure 4.16 have the same pattern with their values being also really similar.

4.4.4 Complete shuffling

The final and most rigorous way of shuffling the data is presented here as complete shuffling. Every GPP value in time and space is replaced by some other GPP value within the data. This is essentially done

(a) (b)

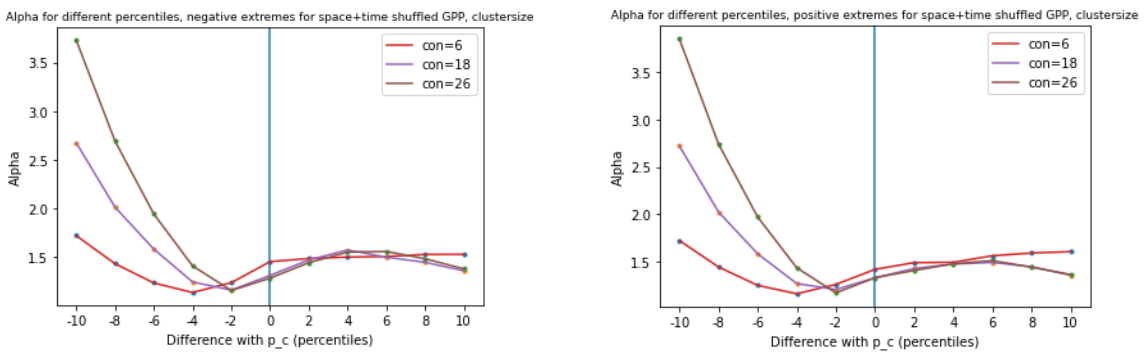

Figure 4.16: The scaling parameter α for cluster sizes of space+time shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 33\%$ for 6-connectivity, $p_c = 16\%$ for 18-connectivity and $p_c = 12\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

by making a long list containing the GPP value for every voxel, then randomly distributing these GPP values across the grid in time and space.

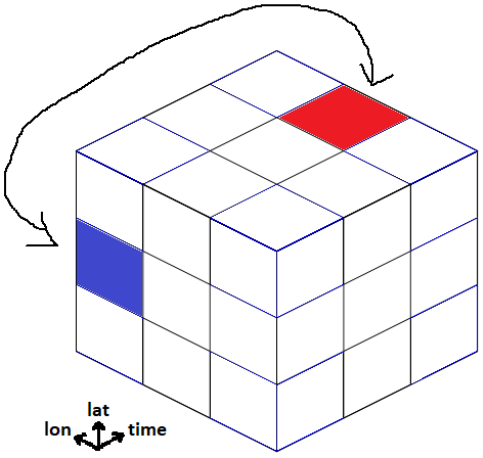

Figure 4.17: A visual representation of complete shuffling of GPP. The red voxel is switched with the blue voxel that is at a different location in time and space.

Size distribution

The size distribution of completely shuffled GPP in figure 4.18a has quite similar characteristics to that of space+time shuffled GPP. There is a small bend at the beginning followed by a power-law region for over 2 orders of magnitude with scaling parameter $\alpha = 1.09$ for both negative and positive extremes. Positive extremes have the largest events, although these are smaller compared to the largest events of all the other cases.

The largest events in terms of cluster sizes are also smaller than those for other cases, as can be seen in figure 4.18b. The power-law regime goes for 2 orders of magnitude and has scaling parameter $\alpha = 1.15$ for both negative and positive extremes.

(a) (b)

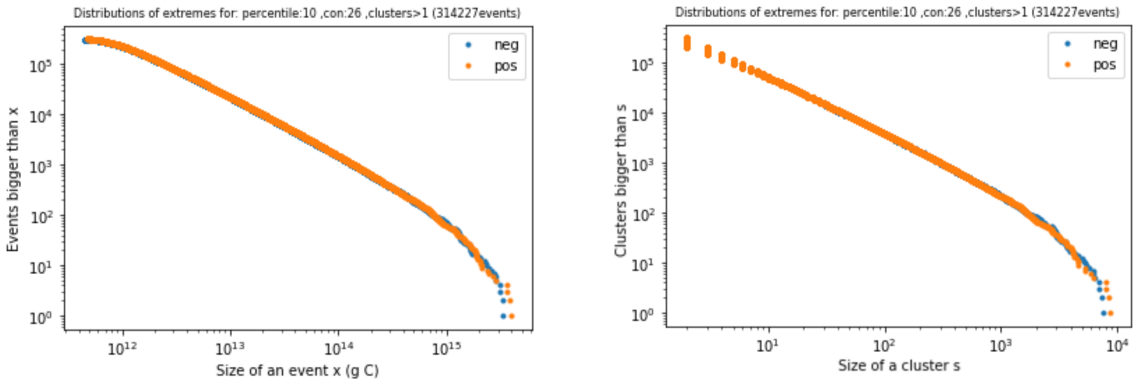

Figure 4.18: **a)** Size distribution of complete shuffled GPP extremes and **b)**cluster sizes for the 10th percentile and 26-connectivity

Table 4.5: Values of the percolation threshold p_c and corresponding exponents τ_{gpp} for the size distribution in completely shuffled GPP, and τ_{cs} for the distribution in cluster sizes.

Con.	6	18	26
p_c	34%	16%	12%
τ_{gpp}	2.20	2.27	2.30
τ_{cs}	2.25	2.28	2.31

Percolation threshold and τ

The values of p_c are almost identical to those for space+time shuffled GPP. The value for 6-connectivity of $p_c = 0.34\%$ is higher, and therefore further from the value predicted by the percolation hypothesis, compared to the space+time shuffled case.

Of all the shuffled data, the values of τ are the closest to the value of $\tau = 2.19$ from percolation theory, especially those of 6-connectivity. The values of τ for 18- and 26-connectivity are slightly higher and similar to those of space+time shuffled. Again, τ for cluster sizes are slightly higher than those for GPP.

Scaling parameter

The plots for α in figure 4.19 look quite similar to those previously seen for space+time shuffled GPP in section 4.4.3. There is a decrease in α until just before the percolation threshold. This is followed by an increase until α finally converges.

The plots for the cluster sizes in figure 4.20 start with quite a sharp decline in α until just before the percolation threshold. After this α rises quite steadily with increasing percentiles.

(a) (b)

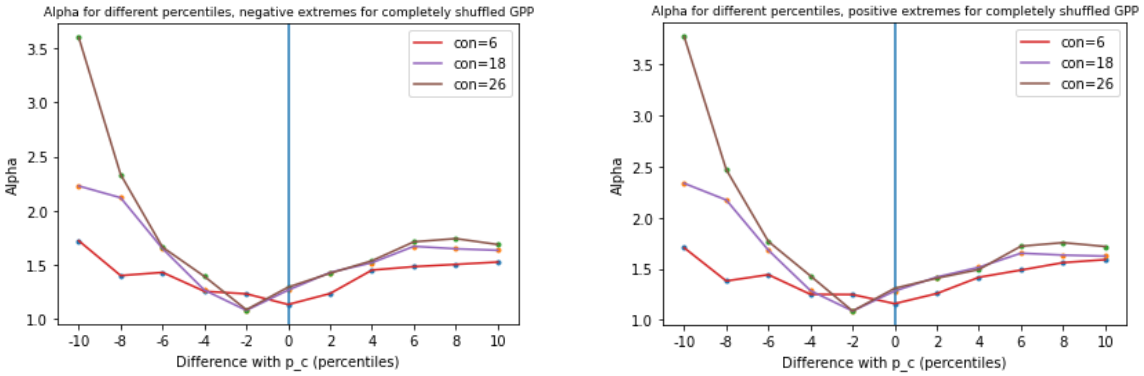

Figure 4.19: The scaling parameter α for completely shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 34\%$ for 6-connectivity, $p_c = 16\%$ for 18-connectivity and $p_c = 12\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

(a) (b)

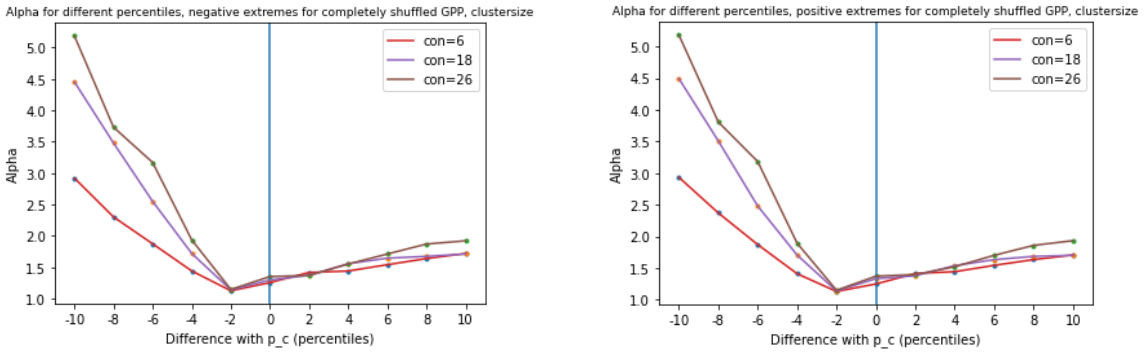

Figure 4.20: The scaling parameter α for cluster sizes of completely shuffled GPP, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 34\%$ for 6-connectivity, $p_c = 16\%$ for 18-connectivity and $p_c = 12\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

4.4.5 Summary and reflection

Correlations in time and space were destroyed in order to be able to get scaling behaviour closer to percolation theory, which assumes uncorrelated data. Randomizing the data by shuffling had clear results on its characteristics. It reduced the size of the largest clusters and increased the total amount of clusters, indicating that the shuffling does indeed has the anticipated effect of destroying the spatiotemporal relations between neighbouring GPP values.

The values of p_c differ in every variation of shuffling. The value of p_c from percolation for 6-connectivity is 31%. Compared to this time shuffling and space shuffling gives values for p_c that are lower than this while space+time shuffling and complete shuffling give values for p_c that are higher. All values of that were found for p_c that were found for 18- connectivity and 26-connectivity were higher than the theoretical values from section 4.1.1, $p_c = 16\%$ and $p_c = 10\%$ respectively. Space+time shuffling and complete shuffling had lower values and thus were closer to the theoretical values of p_c, compared to time shuffling and space shuffling.

All variations of shuffled data had exponents τ that were higher, and therefore closer, to the value of $\tau = 2.19$ from percolation theory compared to unshuffled GPP. Based on this, one could say that the shuffled data gives a result that is closer to the case of percolation theory. Every shuffled case had different values of τ. The most rigorous shuffling methods, space+time shuffling and complete shuffling, had really similar values of τ that were all slightly above the value from percolation theory of $\tau = 2.19$ for 3D-percolation.

It could also be seen that the behaviour of α around the percolation threshold differed between the GPP before and after shuffling. The shapes of the plots for every shuffled case differed as well, although the plots of space+time shuffled GPP and complete shuffled GPP looked very similar. It is hard to give an evaluation on the shape of these plots though, as it is not clear what kind of plots are expected from percolation theory.

These plots for α for space+time shuffled GPP and complete shuffled GPP had some particularly interesting behaviour. In all of these plots there is a point just before p_c after which the slope of the curve of α switches signs. This could be interpreted as critical behaviour and therefore one could hypothesise that this point, in which the slope of the curve of α switches signs, would be at the true value of p_c. In other words, the method that was used for determining p_c might give an overestimation. This would explain why p_c for 6-connectivity was above 31%. Furthermore, it would also explain why the values of τ are higher than in percolation theory. Looking at the plots of α, a lower value of p_c would give a lower corresponding value of α and thus a lower value of τ.

4.5 Synthetic data

In this section synthetic datasets are used where every voxel is independent of each other in time and space. This should result in clusters that are analogous to those of percolation theory. Synthetic datasets are useful as one exactly knows and can control its properties. The results from this dataset can therefore act as a control parameter to compare the shuffled datasets of GPP to. Especially complete shuffled GPP should have really comparable results.

4.5.1 Uniform distribution (land restricted)

A synthetic dataset is generated for the same grid as the one used for GPP, including the restrictions of only using voxels on land. Every voxel is assigned a value generated by an uniform distribution, with the

important characteristic being that the value in all the voxels are independent of eachother. The generated data has values from -0.5 to 0.5, its cumulative size distribution can be seen in figure 4.21.

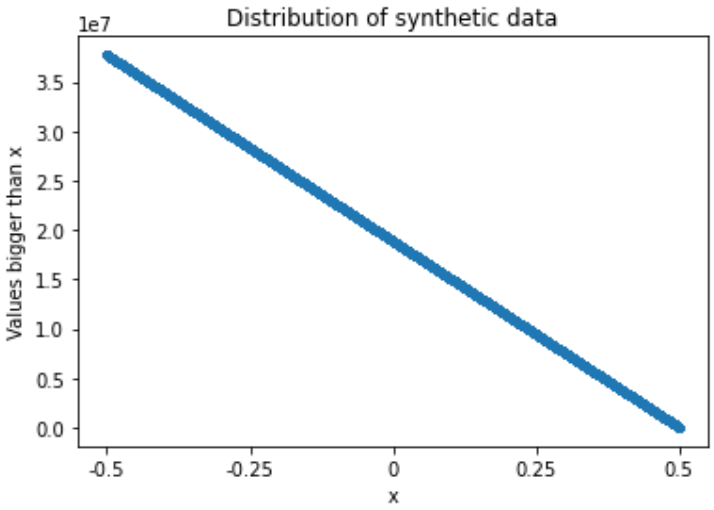

Figure 4.21: Cumulative size distribution of the uniformly distributed synthetic data. x being a value generated by the distribution.

Size distribution

The size distribution of 10th percentile extremes for the synthetic data in figure 4.22a has a power-law region of 2 orders of magnitude with a scaling parameter of $\alpha = 1.10$ and $\alpha = 1.06$ for negative and positive extremes respectively. The power-law regime is followed by a downward bending where the data gets more and more scarce. The total amount of clusters is similar to that of complete shuffled GPP in figure 4.18.

The distribution for cluster sizes in figure 4.22b look almost identical, except for the gaps at the start. The power-law regime that goes over 2 orders of magnitude has a scaling parameter of $\alpha = 1.15$ for both negative and positive extremes, which is identical to that of cluster sizes for complete shuffled GPP. The size of the largest clusters is also almost the same as those for complete shuffled GPP.

(a) (b)

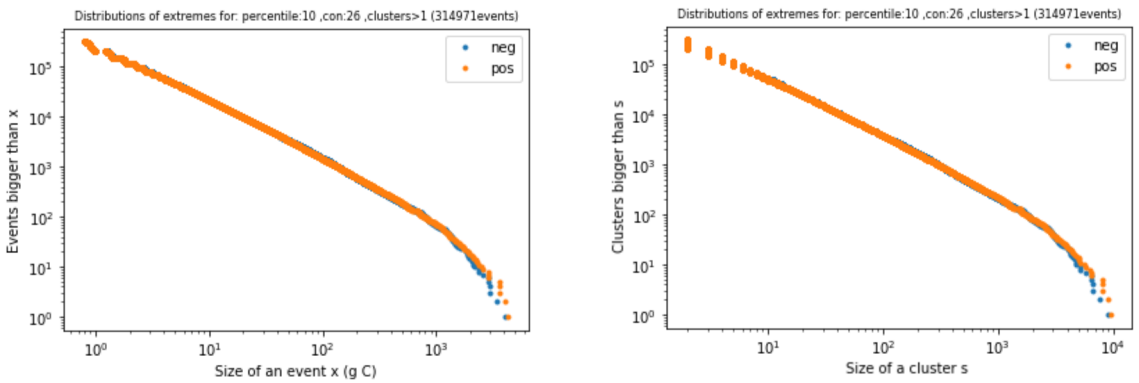

Figure 4.22: **a)** Size distribution of event sizes of synthetic data and **b)**cluster sizes for the 10th percentile and 26-connectivity

Table 4.6: Values of the percolation threshold p_c and corresponding exponents τ_{gpp} for the size distribution in uniform distributed synthetic data, and τ_{cs} for the distribution in cluster sizes.

Con.	6	18	26
p_c	34%	16%	12%
$\tau_{uniform}$	2.21	2.28	2.32
τ_{cs}	2.24	2.29	2.36

Percolation threshold

The values of the percolation threshold p_c in table 4.6 are exactly the same as for completely shuffled GPP. Even the values of τ for 6- and 18-connectivity are really close. This indicates that the shuffling destroyed neighbouring relations enough to make it similar to the uniform synthetic data where all voxels are independent of each other.

Higher connectivities have higher values of τ, and thus are further away from the value of $\tau = 2.19$ from percolation theory.

Scaling parameter

α is plotted for percentiles around the percolation threshold in figure 4.23. There is a decrease in α until $p_c - 2\%$ after which there is a slight increase until α converges to around 1.6 for all connectivities.

(a) (b)

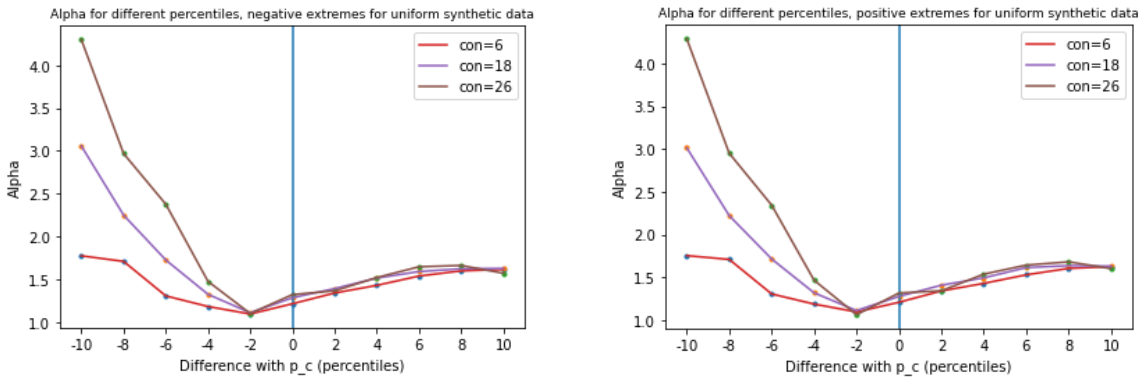

Figure 4.23: The scaling parameter α for event sizes of synthetic data, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 32\%$ for 6-connectivity, $p_c = 15\%$ for 18-connectivity and $p_c = 11\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

The graphs in figure 4.24 for α for the cluster sizes have a similar pattern. The values of α are higher than those of the synthetic data itself. The higher connectivities have higher values of α, especially below the percolation threshold.

4.5.2 Uniform distribution (land and ocean)

Until now, the clustering has been restricted to only voxels on land. However, in percolation theory such a restriction does not apply and every voxel can be part of a cluster. This was named as the third difference

(a) (b)

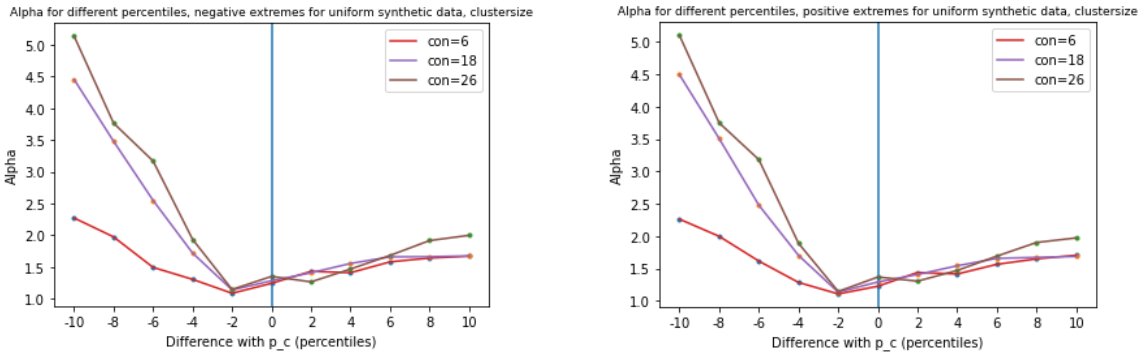

Figure 4.24: The scaling parameter α for cluster sizes of synthetic data, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 32\%$ for 6-connectivity, $p_c = 15\%$ for 18-connectivity and $p_c = 11\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

between clusters in GPP extremes and percolation theory in section 4.1.2. It is however unclear what influence this restriction to land voxels has and how it makes the clustering differ from that of percolation theory.

In this section, a synthetic dataset is generated in the same way as before, except this time every single voxel in the grid, is filled with a value from the same uniform distribution. That means that the previous restriction of only using land-pixels does not apply anymore and also ocean points are covered with values. As an example, in figure 4.25 a map of 10th percentile extremes within a timstep can be seen. The expectation is that this dataset will give power-law scaling behaviour that is the closest to percolation theory from all the datasets that are handled.

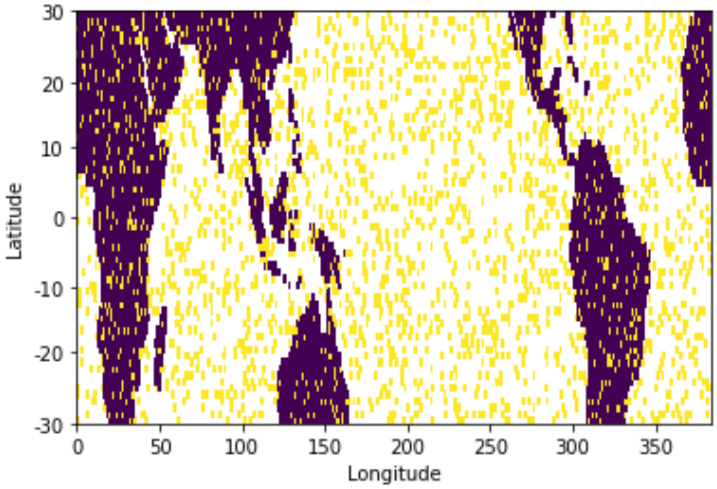

Figure 4.25: An example of the spread of 10th percentile extremes within a timestep. Extremes are marked as yellow while all land is marked as purple. It can be seen that extremes are spread over both land and sea.

Size distribution

The power-law region in the event size distribution in figure 4.26a spans over 2 orders of magnitude and has a scaling parameter of $\alpha = 1.15$ for both negative and positive extremes. After the power-law region

Table 4.7: Values of the percolation threshold p_c and corresponding exponents τ_{gpp} for the size distribution in uniform distributed synthetic data, and τ_{cs} for the distribution in cluster sizes.

Con.	6	18	26
p_c	32%	15%	11%
$\tau_{uniform}$	2.21	2.34	2.30
τ_{cs}	2.21	2.32	2.34

there is a small kink in the curve after which the slope slightly decreases. Towards the end of the curve, at the largest events, there is a downward bending to be seen.

The plots for the cluster sizes in figure 4.26b look similar, except for the gaps in the beginning. The power-law regime that goes over 2 orders of magnitude has a scaling parameter of $\alpha = 1.14$ for both negative and positive extremes.

Compared to the land-restricted synthetic data, the total amount of clusters tripled. This is explained by the fact that including the sea approximately triples the amount of data. It can also be seen that the largest events are an order of magnitude larger compared to the land-restricted synthetic data. It seems like clusters can grow larger when not restricted by continental borders.

(a) (b)

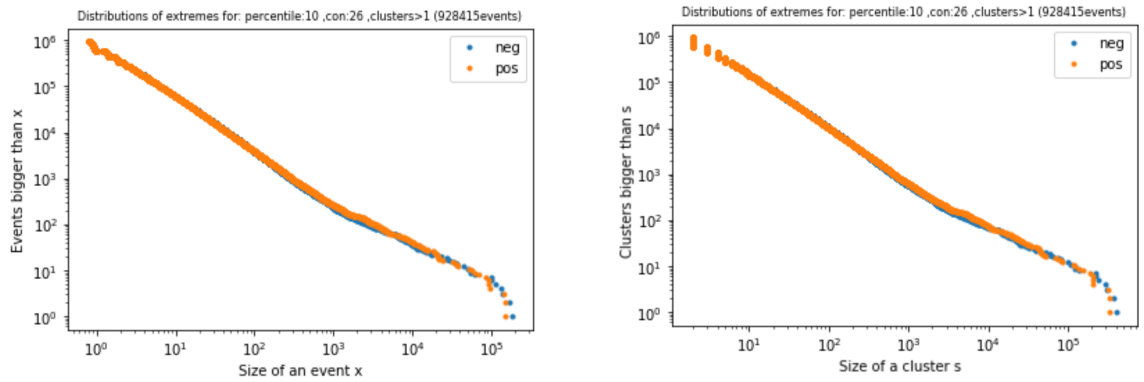

Figure 4.26: **a)** Size distribution of events sizes of synthetic data and **b)**cluster sizes for the 10th percentile and 26-connectivity

Percolation threshold

The values of the percolation threshold p_c are lower compared to the land-restricted synthetic data, as can be see in table 4.7. All values of τ are higher than the value of 2.19 from percolation theory. Similarly to the land-restricted synthetic data, 6-connectivity has lower values of τ than 18- and 26-connectivity.

Scaling parameter

The plots for α in figure 4.27 have similar characteristics as seen before in the land restricted synthetic data, namely a decrease in α until $p_c - 2\%$ followed by an increase in α. However, both this decrease

and increase in α is larger and as a result there are higher values of α further from p_c. Both 18- and 26-connectivity have similar values of α while its values for 6-connectivity are generally lower.

α for cluster sizes in figure 4.28 has the same pattern except that the increase in α after p_c is larger. Thus the values for α are larger for higher percentiles.

(a) (b)

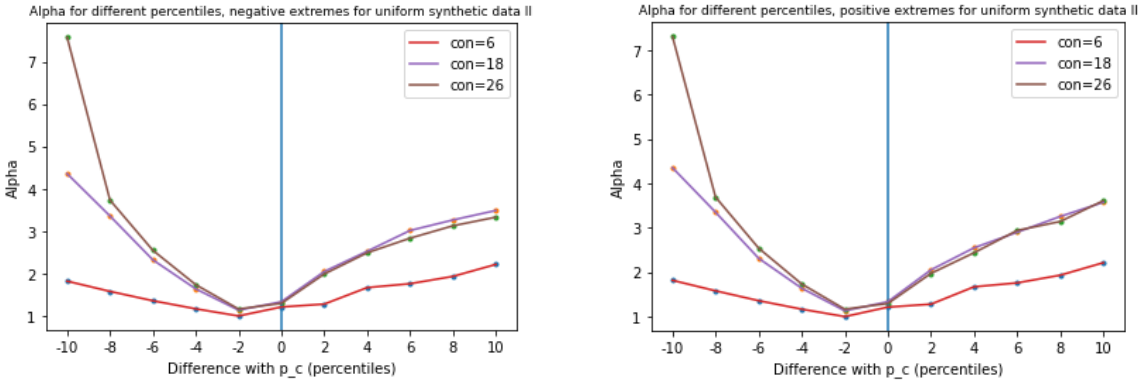

Figure 4.27: The scaling parameter α for event sizes of synthetic data, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 32\%$ for 6-connectivity, $p_c = 15\%$ for 18-connectivity and $p_c = 11\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

(a) (b)

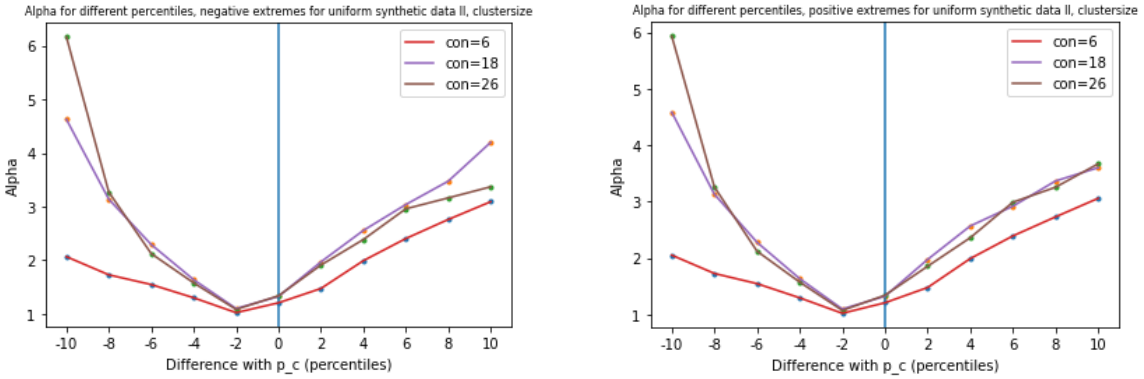

Figure 4.28: The scaling parameter α for cluster sizes of synthetic data, in the percentile range from $p_c - 10\%$ until $p_c + 10\%$. Where $p_c = 32\%$ for 6-connectivity, $p_c = 15\%$ for 18-connectivity and $p_c = 11\%$ for 26-connectivity. **a)** negative extremes **b)** positive extremes

4.5.3 Summary and reflection

The land restricted synthetic data gave similar results to complete shuffled GPP. The size distribution of cluster sizes was almost identical with the scaling parameter α being the exact same. The values of p_c were the same between the cases with τ differing slightly. Thus, complete shuffling did indeed have its desired effect of eliminating all correlations in space and time for GPP.

The second synthetic dataset that was studied had the same uniform distribution but also included voxels in the sea. Its size distribution contained was diferently shaped with much larger events than the land-restricted data. The values of the percolation threshold were slightly lower, thus closer to theoret-

ical values from section 4.1.1, indicating that the continental borders have an restricting effect on the formation of an infinite cluster.

The cluster sizes of the second synthetic dataset should be almost analogous to percolation theory. Yet, all of its values of p_c are higher than the theoretical values from section 4.1.1, and its values of τ are all above the theoretical value of 2.19 as well. Similarly to complete shuffled GPP, there is an overestimation in p_c and therefore τ. The reason and rationale behind this can again be found in the plots for α which provide a similar pattern as those earlier seen for complete shuffled GPP.

4.6 Other related quantities

In the previous sections, power-laws have been found in the size distribution of various types of data. It has become clear that the power-law is not related to the GPP data itself but with the the clustering mechanisms resulting from the particular type of extreme event analysis performed here. Thus, other similar quantities to GPP should also show similar power-law behaviour.

In this section, a brief look will be taken at data for precipitation, sensible heat and latent heat. These quantities are chosen because their extremes are all related to GPP extremes. The data is taken from the same model and handled in the same way as the GPP data. To be able to make a good comparison with GPP, the data is only taken for areas on land.

4.6.1 Precipitation

Extremes in precipitation are related to either droughts or to heavy precipitation events, which effects on GPP were discussed in section 1.3.1 and 1.3.3. Extreme events of precipitation are defined analoguous to the case of GPP by the sum of the precipitation anomalies in grams water within a cluster. Its size distribution is plotted in figure 4.29.

In the double logarithmic plot in 4.29a there is a power-law region followed by a drop-off which seems to occur in two steps. The power-law region covers a little less than 3 orders of magnitude and gives a scaling parameter of $\alpha = 0.61$ and $\alpha = 0.66$ for the negative and positive extremes. Both the width of the power-law region and the value of α are lower than they were for GPP. Nevertheless, there is still clearly power-law scaling behaviour to be seen.

The same size distribution is plotted on a logarithmic y-axis in figure 4.29b. A linear region can be seen at which roughly starts at the point where the drop-off occurs in the double logarithmic plot. This is the same behaviour as could be seen in the plots for GPP.

4.6.2 Temperature (heat fluxes)

Temperature extremes are related to heatwaves and extreme colds and frosts, which effects on GPP were discussed in section 1.3.2 and 1.3.4. As it is quite non-sensical to sum up temperature in the same way as GPP, the quantities sensible heat and latent heat are explored instead. Although heat and temperature are not completely analogous, extremes in heat can be used as a decent approximation to extremes in temperature.

Sensible heat

An extreme event in sensible heat is made up of the total sum of heat anomalies in Joules within a cluster. It size distribution is plotted in figure 4.30.

(a) (b)

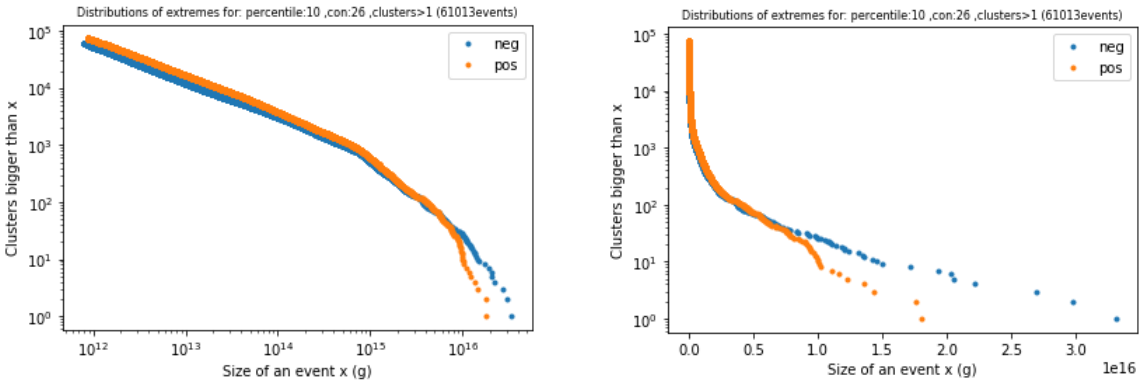

Figure 4.29: Size distribution of precipitation extremes for the 10th percentile and 26-connectivity for **a)** logarithmic x- and y-axis, **b)** logarithmic y-axis

In figure 4.30a it can be seen that the width of the power-law region is almost 4 orders of magnitude, similarly as was the case for GPP. The scaling parameter for the power-law region is $\alpha = 0.74$ and $\alpha = 0.77$ for the negative and positive extremes. Values that are quite close to those of GPP. Towards the end a drop-off occurs which coincides with the beginning of the linear region in the plot with the logarithmic y-axis in figure 4.30b.

(a) (b)

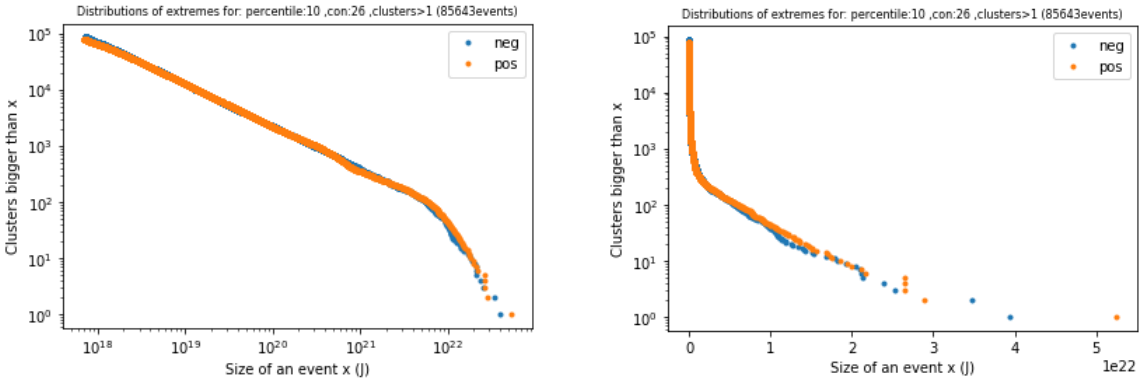

Figure 4.30: Size distribution of sensible heat extremes for the 10th percentile and 26-connectivity for **a)** logarithmic x- and y-axis, **b)** logarithmic y-axis

Latent heat

An extreme event in latent heat is made up in the same way as one was for sensible heat. The size distribution can be seen in figure 4.31.

The double logarithmic plot in figure 4.31 looks quite similar to that of sensible heat. The power-law region is a little bit wider, covering 4 orders of magnitude and giving a scaling parameter of $\alpha = 0.77$ and $\alpha = 0.76$ for the negative and positive extremes. Again, in the plot of the size distribution on a logarithmic y-axis in figure 4.31b, a linear region can be seen starting at the point where the drop-off occurs in the double logarithmic plot.

(a) (b)

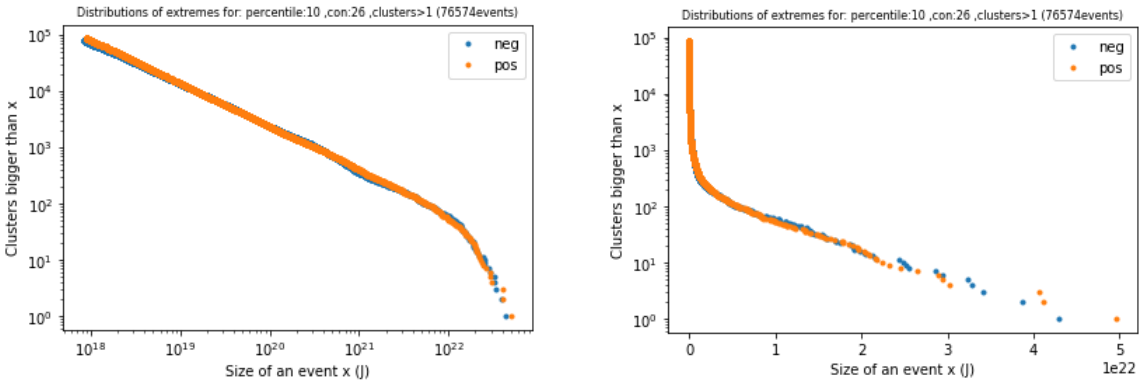

Figure 4.31: Size distribution of latent heat extremes for the 10th percentile and 26-connectivity for **a)** logarithmic x- and y-axis, **b)** logarithmic y-axis

4.6.3 Summary

All three quantities show similar power-law behaviour as GPP. Precipitation has some differences in the shape of the curve of the size distribution and the values of α compared to GPP. Sensible and latent heat behave similar to each other and to GPP, having similar characteristics for its power-law region.

Chapter 5

Discussion and outlook

5.1 Summary and discussion

In this master's thesis I have studied the origin of the power-law behaviour in the size distribution of extreme events of GPP that was found by Zscheischler et al. (2013)[1].

I replicated the methodology of Zscheischler et al. using simulation data instead of observational data, and found a similar power-law. The width of the power-law region in my plots are larger than in the plots of Zscheischler et al., giving more reliable scaling behaviour. This difference is caused by the larger amount of data that I had access to using simulation data from CMIP6, compared to the observational data of Zscheischler et al. 10th percentiles extremes have a clearer power law region compared to lower percentiles. Exponential behaviour is seen in plots with only a logarithmic y-axis, which is in line with equation 1.5 from section 1.4.1 about percolation theory.

A hypothesis was made that the origin of the power-law can be found in the distribution of the GPP data itself. However, both the size distributions of GPP and GPP anomalies did not display any power-law behaviour, so this hypothesis was rejected.

Since the power-law did not come from the GPP data, its origin may be instead in the clustering and the mechanisms behind it. Thus, a new hypothesis was formed centered around percolation theory, based on the fact that both GPP extremes and percolation theory are concerned with clusters made out of a certain fraction of the data. This certain fraction is made up by "percentiles" for GPP extremes and "probability" in percolation theory. Moreover, the exponent α for the power-law in the size distribution of GPP is related to the exponent τ describing cluster sizes in percolation theory by the relation $\tau = \alpha + 1$. The values of the percolation threshold p_c and τ of GPP differ from the values from percolation theory. This can be seen in an overview of all values that were found for p_c in table 5.1 and for τ in table 5.2. The difference in p_c and τ is caused by differences between clustering for GPP extremes and percolation theory, namely concerning the differing value of the voxels, correlations in time and space, and the restriction of GPP values to land.

Next, the cluster sizes of GPP extreme events were looked at to eliminate the difference that GPP values within the clusters caused with percolation theory. The obtained values of τ are closer to that of percolation theory.

Next, the GPP data was randomized by a process called shuffling in order to destroy correlations in time and space and to investigate whether this creates behaviour closer to percolation theory, which assumes uncorrelated data. Shuffling was done in time, space, space+time and completely. More rigorous shuffling results in values of p_c of τ that are closer to percolation theory. Based on the values of α around p_c, it was theorized that some of the obtained values of p_c and τ were an overestimation.

Then, synthetic data generated using an uniform distribution was used where every voxel was independent of each other. One dataset had the same restrictions to land as GPP had, while the other dataset had data points in both land and sea. The synthetic dataset that was restricted to land has similar results to complete shuffled GPP regarding its size distribution and its values of p_c and τ. Removing the restriction to land for the synthetic dataset changes the shape of its size distribution and its values of p_c and τ. The values of p_c got slightly lower, thus closer to theoretical values from section 4.1.1, indicating that the continental borders have an restricting effect on the formation of an infinite cluster.

It could be seen that altering the GPP data to eliminate the differences to percolation theory did indeed result in scaling behaviour that is closer to percolation theory. Indeed, it can be seen that for shuffled and for synthetic data the values of both p_c in table 5.1 and τ in table 5.2 are closer to values that are predicted by percolation theory. The theoretical values for p_c and τ could however not be achieved. One important reason for this can be found in the uncertainty and errors in the method of determination of both p_c and τ. Still, both the complete shuffled data and the synthetic data give values of p_c and τ close to those from percolation theory, indicating similar power-law scaling behaviour. This can be interpreted as that the clustering mechanisms behind extreme event analysis are similar to the clustering in percolation theory. Thus, one can conclude that percolation theory is a reasonable explanation behind the power-law in GPP extremes.

Lastly, the size distribution of other quantities related to GPP were taken a look at. Precipitation, sensible heat and latent heat all had power-law behaviour similar to GPP, proving that the power-law that was found by Zscheischler et al. is not exclusive to GPP and can be found in other quantities as well.

The main research question that was posed in the beginning "What is the origin of the power-law behaviour in the size distribution of GPP extreme events?" can now be answered: The origin of the power-law behaviour does not depend on GPP, in general it does not depend on the data itself but on the clustering mechanisms underlying percolation theory.

5.2 Outlook

There are several things that can be improved in this study. These can serve as a recommendation for further studies regarding the subject of percolation in GPP (and other spatiotemporal fields).

For example, I was working on a proof of why percolation behaviour can be seen in event sizes of GPP, despite the fact that GPP has different values for different voxels, while in percolation theory every voxel essentially has the same value. This proof was based on the fact that for large enough clusters, the mean GPP value of a voxel in an event would be the same as the mean GPP value of all voxels containing an extreme value. Essentially, this would mean that one would end up in a situation where the different GPP values would not cause a difference between event sizes for clusters of the same size. Thus, one would end up with scaling behaviour for event sizes that is similar to cluster sizes. Unfortunately, this proof could not be further explored due to restrictions in time.

Next are the methods of determination of α and p_c which are two key parameters in this study. The main improvement for the determination of α lies in a better choice of the power-law region. More reliable methods could be sought out for the determination of both the starting and the ending point of the linear region, which would result in better values for α.

Based on the results, it seemed like the method of determination of p_c resulted in an overestimation of its value. The reason is that an infinite cluster is only searched for in the dimension of time. An infinite cluster can however also occur in space before it occurs in time, but this is not detected in my method. In addition, the precision in the value of p_c, which is only determined to 2 decimals due to time constraints,

Table 5.1: An overview of p_c values. p_{c6} stands for values for 6-connectivity, p_{c18} stands for values for 18-connectivity and p_{c26} stands for values for 26-connectivity

Data	p_{c6}	p_{c18}	p_{c26}
Percolation theory	31%	14%	10%
Original GPP	26%	20%	20%
Time shuffled	27%	21%	19%
Space shuffled	25%	16%	14%
Space+time shuffled	33%	16%	12%
Complete shuffled	34%	16%	12%
Uniform (land)	34%	16%	12%
Uniform (land+ocean)	32%	15%	11%

is sub-optimal. If a better value of p_c is desired, one should use a better method of finding an infinite cluster and more precisely determine the resulting value of p_c.

As determination of τ is affected by both α and p_c, an improvement in the method of determination of both would result in better values of τ as well.

Table 5.2: An overview of values of τ. τ_{gpp6} stands for values of GPP for 6-connectivity, τ_{cs6} stands for values of cluster sizes for 6-connectivity. Analogously τ_{gpp18}, τ_{cs18}, τ_{gpp26} and τ_{cs26} stands for its respective values of 18- and 26-connectivity.

Data	τ_{gpp6}	τ_{cs6}	τ_{gpp18}	τ_{cs18}	τ_{gpp26}	τ_{cs26}
Percolation theory	2.19	2.19	2.19	2.19	2.19	2.19
Original GPP	1.66	1.96	1.72	1.84	1.71	1.83
Time shuffled	1.87	1.98	1.81	1.89	1.82	1.88
Space shuffled	2.16	2.22	2.04	2.28	2.06	2.28
Space+time shuffled	2.29	2.31	2.31	2.32	2.31	2.30
Complete shuffled	2.20	2.25	2.27	2.28	2.30	2.31
Uniform (land)	2.21	2.24	2.28	2.29	2.30	2.36
Uniform (land+ocean)	2.21	2.21	2.34	2.32	2.30	2.34

Appendix A

A.1 Proof of the relation $\tau = \alpha + 1$

The power-law exponents of the GPP size distribution α and of cluster sizes in percolation theory τ are related by the relation $\tau = \alpha + 1$. This will be proven in this section.

The y-axis of the size distribution of GPP extreme events, I will call this quantity y_s, describes the value of events or clusters that are larger than a certain size s. In other words, it is the sum of the number of clusters n_s, from s to the largest cluster s_{max}:

$$y_s = \sum_{s}^{s_{max}} n_s \tag{A.1}$$

If the biggest clusters are sufficiently large:

$$y_s = \lim_{s_{max}\to\infty} \sum_{s}^{s_{max}} n_s \simeq \int_{s}^{s_{max}} n_s ds \tag{A.2}$$

filling in equation 1.3:

$$y_s = \int_{s}^{s_{max}} s^{-\tau} ds = s^{-\tau+1} \tag{A.3}$$

combined with the scaling equation of y_s:

$$y_s = s^{-\alpha} \tag{A.4}$$

this will finally gives the relation between τ and α:

$$y_s = s^{-\alpha} = s^{-\tau+1} \tag{A.5}$$

from which it can seen that $\tau = \alpha + 1$.

A.2 Piecewise linear functions (pwlf)

Piecewise linear functios (pwlf) is used to determine the starting point of the power-law region in the size distribution of extreme events. Pwlf is a python library that can be used to fit piecewise linear functions [42]. So called ”breakpoints” that mark the beginning or end of a linear region are found by using global optimization for a specified number of linear segments.

First, the desired number of breakpoints is specified. The first and last breakpoints are assumed to be at the beginning and the end of the curve. The rest of the breakpoints are chosen so that the sum-of-squares of the residuals (SSR)

$$SSR = e^T e \tag{A.6}$$

is minimized. The squared norm of e is taken here, e being the residual vector that is the difference between the fitted continuous piecewise linear model and the original dataset.

A.3 Python function numpy.polyfit

Numpy.polyfit is used to determine the scaling parameter α in the size distribution of extreme events by performing a linear fit on the power-law region.

Numpy.polyfit is a python function that fits a polynomial $p(x) = p[0] * x^deg + ... + p[deg]$ of degree deg to points (x,y) [43]. For linear fitting $deg = 1$ is used. The coefficients of the fit are based on minimizing the least squares error.

A.4 Fisher-Yates shuffling

Fisher-Yates shuffling is an algorithm for generating random unbiased permutations of a finite sequence. It is named after Ronald Fisher and Frank Yates who first described it in 1938 [44]. A modern version of this algorithm, designed for computer use, was introduced by Durstenfeld in 1964 [45]. This modern version of the algorithm reduces the time complexity, and is therefore the version that is used in this thesis.

An example of the modern version of the shuffling method can be seen in figure A.1. In the first iteration, the last element is ”struck”. This element is then swapped with one randomly selected element. If this randomly selected element happens to be the same, no swap will take. For the next iterations, the last ”un-struck” element is swapped with one randomly selected ”un-struck” element. This is repeated until every element is ”struck” resulting in a random permutation of the originals elements.

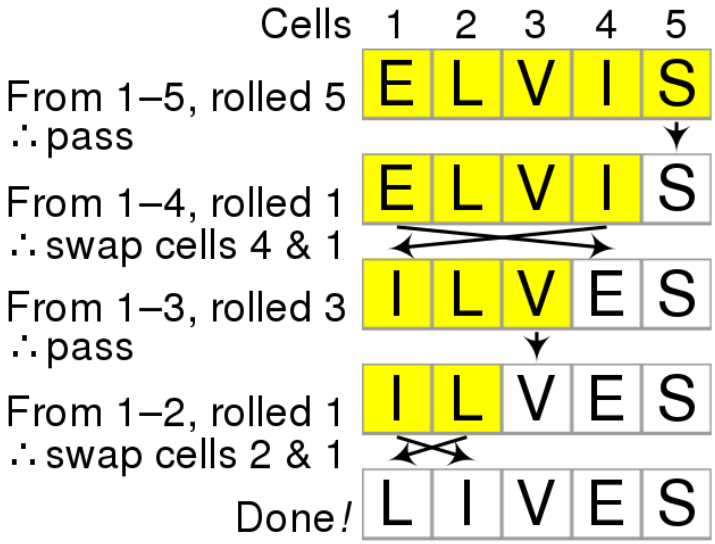

Figure A.1: An explanatory image of the modern version of Fisher-Yates shuffling performed on five letters. Reproduced from [6]

Bibliography

[1] Jakob Zscheischler, Miguel D Mahecha, Stefan Harmeling, and Markus Reichstein. Detection and attribution of large spatiotemporal extreme events in earth observation data. *Ecological Informatics*, 15:66–73, 2013.

[2] Jakob Zscheischler, Miguel D Mahecha, Jannis Von Buttlar, Stefan Harmeling, Martin Jung, Anja Rammig, James T Randerson, Bernhard Schölkopf, Sonia I Seneviratne, Enrico Tomelleri, et al. A few extreme events dominate global interannual variability in gross primary production. *Environmental Research Letters*, 9(3):035001, 2014.

[3] Percolation model and controllability. `https://www.slideshare.net/MohammadrezaDehghani1/percolation-model-and-controllability`.

[4] Compute clique statistics. `https://brainvisa.info/axon/en/processes/AtlasComputeCliqueFromLabels.html`.

[5] J. Zscheischler. A global analysis of extreme events and consequences for the terrestrial carbon cycle. 2014.

[6] Fisher–yates shuffle. `https://www.wikiwand.com/en/Fisher-Yates_shuffle`.

[7] Boris I Shklovskii and Alex L Efros. Percolation theory. In *Electronic Properties of Doped Semiconductors*, page 138. Springer, 1984.

[8] Jakob Zscheischler, Markus Reichstein, S Harmeling, A Rammig, Enrico Tomelleri, and Miguel D Mahecha. Extreme events in gross primary production: a characterization across continents. *Biogeosciences*, 11(11):2909–2924, 2014.

[9] Markus Reichstein, Michael Bahn, Philippe Ciais, Dorothea Frank, Miguel D Mahecha, Sonia I Seneviratne, Jakob Zscheischler, Christian Beer, Nina Buchmann, David C Frank, et al. Climate extremes and the carbon cycle. *Nature*, 500(7462):287–295, 2013.

[10] David Francis and Henry Hengeveld. *Extreme weather and climate change*. Environment Canada Ontario, 1998.

[11] John A Arnone Iii, Paul SJ Verburg, Dale W Johnson, Jessica D Larsen, Richard L Jasoni, Annmarie J Lucchesi, Candace M Batts, Christopher von Nagy, William G Coulombe, David E Schorran, et al. Prolonged suppression of ecosystem carbon dioxide uptake after an anomalously warm year. *Nature*, 455(7211):383–386, 2008.

[12] David D Breshears, Neil S Cobb, Paul M Rich, Kevin P Price, Craig D Allen, Randy G Balice, William H Romme, Jude H Kastens, M Lisa Floyd, Jayne Belnap, et al. Regional vegetation die-off in response to global-change-type drought. *Proceedings of the National Academy of Sciences*, 102(42):15144–15148, 2005.

[13] Dietrich Stauffer, Amnon Aharony, and Sidney Redner. Introduction to percolation theory. *Physics Today*, 46(4):64, 1993.

[14] Martin Jung, Markus Reichstein, Hank A Margolis, Alessandro Cescatti, Andrew D Richardson, M Altaf Arain, Almut Arneth, Christian Bernhofer, Damien Bonal, Jiquan Chen, et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences*, 116(G3), 2011.

[15] Dorothea Frank, Markus Reichstein, Michael Bahn, Kirsten Thonicke, David Frank, Miguel D Mahecha, Pete Smith, Marijn Van der Velde, Sara Vicca, Flurin Babst, et al. Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts. *Global change biology*, 21(8):2861–2880, 2015.

[16] Shilong Piao, Xinping Zhang, Anping Chen, Qiang Liu, Xu Lian, Xuhui Wang, Shushi Peng, and Xiuchen Wu. The impacts of climate extremes on the terrestrial carbon cycle: A review. *Science China Earth Sciences*, 62(10):1551–1563, 2019.

[17] Sebastian Sippel, Markus Reichstein, Xuanlong Ma, Miguel D Mahecha, Holger Lange, Milan Flach, and Dorothea Frank. Drought, heat, and the carbon cycle: a review. *Current Climate Change Reports*, 4(3):266–286, 2018.

[18] Craig D Allen, David D Breshears, and Nate G McDowell. On underestimation of global vulnerability to tree mortality and forest die-off from hotter drought in the anthropocene. *Ecosphere*, 6(8):1–55, 2015.

[19] Ph Ciais, M Reichstein, Nicolas Viovy, André Granier, Jérôme Ogée, Vincent Allard, Marc Aubinet, Nina Buchmann, Chr Bernhofer, Arnaud Carrara, et al. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437(7058):529–533, 2005.

[20] Garrett P Marino, Dale P Kaiser, Lianhong Gu, and Daniel M Ricciuto. Reconstruction of false spring occurrences over the southeastern united states, 1901–2007: an increasing risk of spring freeze damage? *Environmental Research Letters*, 6(2):024015, 2011.

[21] Jeffrey A Hicke, Craig D Allen, Ankur R Desai, Michael C Dietze, Ronald J Hall, Edward H Hogg, Daniel M Kashian, David Moore, Kenneth F Raffa, Rona N Sturrock, et al. Effects of biotic disturbances on forest carbon cycling in the u nited s tates and c anada. *Global Change Biology*, 18(1):7–34, 2012.

[22] MJB Zeppel, Janet V Wilks, and James D Lewis. Impacts of extreme precipitation and seasonal changes in precipitation on plants. *Biogeosciences*, 11(11):3083–3093, 2014.

[23] Chuixiang Yi, Elise Pendall, and Philippe Ciais. Focus on extreme events and the carbon cycle. *Environmental Research Letters*, 10(7), 2015.

[24] Bruce D Malamud, Gleb Morein, and Donald L Turcotte. Forest fires: an example of self-organized critical behavior. *Science*, 281(5384):1840–1842, 1998.

[25] Stijn Hantson, Salvador Pueyo, and Emilio Chuvieco. Global fire size distribution: from power law to log-normal. *International journal of wildland fire*, 25(4):403–412, 2016.

[26] Ole Peters, Christopher Hertlein, and Kim Christensen. A complexity view of rainfall. *Physical review letters*, 88(1):018701, 2001.

[27] Ole Peters and Kim Christensen. Rain: Relaxations in the sky. *Physical Review E*, 66(3):036120, 2002.

[28] Jan F Eichner, Eva Koscielny-Bunde, Armin Bunde, Shlomo Havlin, and H-J Schellnhuber. Power-law persistence and trends in the atmosphere: A detailed study of long temperature records. *Physical Review E*, 68(4):046133, 2003.

[29] Álvaro Corral and Álvaro González. Power law size distributions in geoscience revisited. *Earth and Space Science*, 6(5):673–697, 2019.

[30] Tomer Kalisky and Reuven Cohen. Width of percolation transition in complex networks. *Physical Review E*, 73(3):035101, 2006.

[31] Abbas Ali Saberi. Percolation description of the global topography of earth and the moon. *Physical review letters*, 110(17):178501, 2013.

[32] Franziska Taubert, Rico Fischer, Jürgen Groeneveld, Sebastian Lehmann, Michael S Müller, Edna Rödig, Thorsten Wiegand, and Andreas Huth. Global patterns of tropical forest fragmentation. *Nature*, 554(7693):519–522, 2018.

[33] Abbas Ali Saberi and Horr Dashti-Naserabadi. Three-dimensional ising model, percolation theory and conformal invariance. *EPL (Europhysics letters)*, 92(6):67005, 2011.

[34] MF Sykes and JW Essam. Critical percolation probabilities by series methods. *Physical Review*, 133(1A):A310, 1964.

[35] Naeem Jan and Dietrich Stauffer. Random site percolation in three dimensions. *International Journal of Modern Physics C*, 9(02):341–347, 1998.

[36] Dietrich Stauffer. Scaling theory of percolation clusters. *Physics reports*, 54(1):1–74, 1979.

[37] Paul Sotta and Didier Long. The crossover from 2d to 3d percolation: theory and numerical simulations. *The European Physical Journal E*, 11(4):375–388, 2003.

[38] Johann Jungclaus, Matthias Bittner, Karl-Hermann Wieners, Fabian Wachsmann, Martin Schupfner, Stephanie Legutke, Marco Giorgetta, Christian Reick, Veronika Gayler, Helmuth Haak, Philipp de Vrese, Thomas Raddatz, Monika Esch, Thorsten Mauritsen, Jin-Song von Storch, Jörg Behrens, Victor Brovkin, Martin Claussen, Traute Crueger, Irina Fast, Stephanie Fiedler, Stefan Hagemann, Cathy Hohenegger, Thomas Jahns, Silvia Kloster, Stefan Kinne, Gitta Lasslop, Luis Kornblueh, Jochem Marotzke, Daniela Matei, Katharina Meraner, Uwe Mikolajewicz, Kameswarrao Modali, Wolfgang Müller, Julia Nabel, Dirk Notz, Karsten Peters-von Gehlen, Robert Pincus, Holger Pohlmann, Julia Pongratz, Sebastian Rast, Hauke Schmidt, Reiner Schnur, Uwe

Schulzweida, Katharina Six, Bjorn Stevens, Aiko Voigt, and Erich Roeckner. Mpi-m mpi-esm1.2-hr model output prepared for cmip6 cmip picontrol, 2019.

[39] Wolfgang A Müller, Johann H Jungclaus, Thorsten Mauritsen, Johanna Baehr, Matthias Bittner, R Budich, Felix Bunzel, Monika Esch, Rohit Ghosh, Helmut Haak, et al. A higher-resolution version of the max planck institute earth system model (mpi-esm1. 2-hr). *Journal of Advances in Modeling Earth Systems*, 10(7):1383–1413, 2018.

[40] Miguel D Mahecha, Lina M Fürst, Nadine Gobron, and Holger Lange. Identifying multiple spatiotemporal patterns: A refined view on terrestrial photosynthetic activity. *Pattern Recognition Letters*, 31(14):2309–2317, 2010.

[41] Ronald Aylmer Fisher and Frank Yates. *Statistical tables for biological, agricultural and medical research*. Hafner Publishing Company, 1953.

[42] Charles F Jekel and Gerhard Venter. pwlf: a python library for fitting 1d continuous piecewise linear functions. *URL: https://github. com/cjekel/piecewise_linear_fit_py*, 2019.

[43] numpy.polyfit. `https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html#numpy-polyfit`.

[44] Ronald Aylmer Fisher and Frank Yates. *Statistical tables for biological, agricultural and medical research*. London: Oliver & Boyd, 1948 [1938].

[45] Richard Durstenfeld. Algorithm 235: random permutation. *Communications of the ACM*, 7(7):420, 1964.