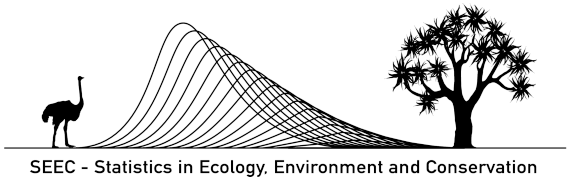
Changes in Rainfall Seasonality in the Western Cape, South Africa: An Exploration of Methods for Determining the Start and End of the Rainfall Season

University of Cape Town - Masters in Statistical Sciences

2018 - 2019

Peter Ivey

Supervisor: Birgit Erni



Contents

1. Introduction

The aim of this thesis is to detect and analyse changes in seasonality in rainfall for various groups of weather stations in the Western Cape area. Weather stations with similar seasonal patterns are firstly grouped together using certain clustering algorithms. The start and end of the rainfall season dates for the different groups of weather stations are estimated and then compared over time to determine whether there have been any changes. Once these start and end of season dates have been estimated, the length of the rainfall season is estimated and compared over time.

Studies have been performed globally and over southern Africa attempting to analyse rainfall patterns and changes. However, rainfall is the most unstable climate variable in terms of time and space and thus, it is really difficult to predict (Yaman, 2018). Most studies have pointed toward an increase of extreme events on both sides of the scale i.e. more intense flooding and more severe drought being experienced. Some places are also starting to experience more rainfall than before whilst other places are starting to experience more drought. The impacts of these rainfall changes are already being experienced with many areas being forced to adapt to the new conditions.

Many better decisions can be made with a better understanding of how rainfall seasons are changing. In the agricultural industry, better informed decisions about when the rainfall season is likely to start and end can result in more optimal yield from crops. Changes in rainfall can also affect the type of crops that should be planted. Farmers will also be able to better prepare for drought seasons if they are better informed as to when these drought periods will likely occur. In terms of disaster risk management, the more that is known about rainfall patterns, the better prepared regions can be for the inevitable increase in extreme events. Cities can put in better systems now in order to deal with potential future crises. Cape Town is an example of a city that could have possibly been better prepared for the current drought crisis if there was a better understanding of rainfall trends. Hopefully in the future, with more accurate information about rainfall, it can rather be an active process than a reactionary process to the current climate conditions.

1.1 The Western Cape and Rainfall

The Western Cape can be found at the most southern tip of South Africa and is home to one of the country's major cities. Cape Town, a city that continues to grow, is one of the hot spot tourist destinations of the world. Large-scale irrigated agriculture is at the forefront of economic growth due to fruit exportation and wine production. The region relies on winter rainfall to supply water throughout the year and so, if not enough water is captured and stored throughout winter, drought is likely to be experienced. Even though the region is drought prone, not much is known about rainfall variability and how the seasons may have shifted over the years.

Water restrictions are frequently employed to reduce the consumption of fresh water. Drought can lead to, and has in the past lead to severe agricultural losses causing job loss, resulting in a negative impact on the economy. This makes the study of the climate, in particular, rainfall variability, of utmost importance. With improved knowledge about the climate and rainfall patterns, more optimal decisions can be made regarding agricultural production and more strategic water usage plans can be implemented.

1.2 Climate Structure of the Western Cape

The Western Cape can be divided into three separate climate regions (Van Niekerk et. al., 2011). These three different regions are as a result of the Cape Fold Belt - a mountain range that forms an 'L' shape throughout the Western Cape. The three regions all have their own unique climatic conditions based on various geographical features. The first region - the Mediterranean region - is found on the southwest coast of the Western Cape. This region, typical of a Mediterranean climate, experiences a dry summer and winter rainfall. These conditions are a result of the combination of the cold Bengeula current on the west coast as well as the northward movement of the South Atlantic High Pressure Cell during winter, allowing rain-bearing mid-latitude cyclones to pass over the southwest coast of SA (du Plesssis et. al., 2017).

Toward the southeast of the Western Cape, the South Coast region is found. The defining point for the start of this region is at Cape Agulhas - the meeting point of the cold Atlantic Ocean and warm Indian Ocean. The region extends eastward and inland until the start of the Cape Fold Belt. As a result of the warm Indian Ocean, this region experiences all year rainfall. This region is also affected by the Southern Oscillation which influences the amount of rainfall experienced along the east coast.

The third region - the Karoo - is found inland on the plateau of the Cape Fold Belt. This region has no influence from the ocean or mid-latitude cyclones. It experiences all year rainfall with maximum rainfall occurring during summer accompanied by thunder storms (Van Niekerk et. al., 2011).

1.3 Summary

This first chapter highlighted the aim of the thesis and why more research needs to be done into rainfall seasonality. In chapter 2, certain literature is reviewed looking at changes in rainfall seasonality in different parts of the world, various methods to detect seasonality change, what other climate variables have been assessed and lastly, the potential environmental impacts of changing seasons. Chapter 3 describes the methodology that is used to cluster the weather stations, estimate the start and end of season and estimate the length of season. Chapter 4 highlights the obtained results and chapter 5 explains the final conclusions.

2. Literature Review

Water is vital to life on earth and as the population continues to expand, its importance ever increases (Yaman, 2014). As a result, much research has been done to try model rainfall and how it has changed over time. Rainfall tends to fall seasonally in most regions and it is important to know when these seasons begin and end and how they are lengthening or shortening. In this chapter, progress into the research of seasonality is tracked, highlighting seminal academic work.

2.1 A change in rainfall trend on a global and local scale

Many rainfall patterns appear to be in the process of gradually changing. Some areas are receiving more rainfall and are having a longer lasting rainfall season whilst other regions are seeing less rainfall as well as a decrease in the length of the rainfall season. There also appears to be a shift toward more extreme rainfall events being experienced globally (Pohl et. al., 2017). Even though it is often difficult to see statistically significant trends, changes in seasonality are evident.

Evidence for the change of many climatological patterns, including the change of rainfall trends, are widely present (du Plessis et. al., 2017). These changes are as a result of possible natural shift in the climate as well as the increased effect of humans on 'climate change'. Kruger (et, al., 2017) conclude that since the concept of 'global warming' and now 'climate change' have gained more prominence, many studies have been performed analyzing certain climate trends - particularly the changes in temperature and precipitation. The purpose of many of these studies is to assess how climate change is having impacts on both local and global scales. Even though many studies have been performed globally, within the more developing countries, due to lower data quality, there have been fewer studies and so, less is known about rainfall within these countries (New et al., 2006). South Africa is one exception within the southern African region that has a substantial amount of rainfall data and so analysis of this data is important so as to inform South Africa of how rainfall is changing, whilst at the same time, supplying the surrounding regions with some idea of the changing rainfall patterns (Kruger, 2006).

Considering the broader area of southern Africa, there is evidence that rainfall patterns are gradually changing over time. It can be fairly difficult to detect whether these observed changes are of any significance or whether they are just part of a larger cycle. A study performed on rainfall data covering various countries over southern Africa looking at trends in daily climate extremes over southern and west Africa points toward an increase in regional average rainfall intensity as well as an increase in dry spell length (New et. al., 2006). This suggests that there is a decrease in total rainfall and at the same time, days of precipitation will be fewer, but they will be more intense. Similar studies for the same region echo this conclusion. In a study performed by Pohl et. al. (2017) over the southern Africa region, it was concluded that there would be a decrease in the number of rainy days together with more intense rainfall on those rainy days. The increase in more extreme events is attributed to the increasing effect of 'climate change'. It is believed that with the continue release of greenhouse gases, the hydrological cycle will continue to intensify explaining why many more extreme events are being experienced globally.

On a more international scale, many studies have been performed, particularly in the more developed countries. In the United States of America, Dourte et. al. (2015) analysed rainfall data in the southeastern US in order to detect any changes in intensity and seasonality. A significant increase in the number of extreme rainfall days and a decrease in the number of low intensity days was noted as well as evidence of increased variability in spring and summer rainfall patterns. Murphy et. al. (2006) performed analysis on certain climate variables over southeastern Australia (SEA) - where most of the population is concentrated. Through the period of 1996 to 2007, SEA has suffered from very low rainfall. Over this decade, rainfall was just over 14% below the climatological mean for the previous three decades. Significantly drier autumns are the main reason for the decrease in rainfall indicating some seasonal shift.

2.2 Seasonal Rainfall in South Africa

In South Africa, predominantly summer rainfall but also winter rainfall is experienced within different regions of the country. These areas are largely influenced by the oceans, various other climatic conditions and the presence of mid-latitude cyclones along the south coast. The southern tip of South Africa is affected by cold fronts that form throughout the year, but typically only hit the south coast of South Africa during the winter months (May to September) when the South Atlantic High Pressure cell has shifted further north. Thus, winter rainfall is experienced along the south coast in the Western Cape. This region of South Africa is one of the few that experiences winter rainfall. This region is also influenced by the Antarctic Oscillation - the main pattern of tropospheric circulation variability below 20 degrees south (Reason et. al., 2005). Typically, a wet winter appears to be associated with the negative phase of the Antarctic Oscillation and vice versa for a dry winter.

2.2.1 The East Coast of South Africa

The east coast of South Africa is largely affected by the warm Agulhas Current that runs down the coast. Due to the warm sea temperature and tropical conditions, winds from the south bring summer rainfall during the months of November to March. Rainfall seasonality along the east coast also changes as a result of the Southern Oscillation (van Heerden, 1994). The Southern Oscillation consists of the El Nino Southern Oscillation (ENSO) as well as the La Nina Southern Oscillation (LNSO). The ENSO is associated with the warming of the tropical pacific and the LNSO is associated with the cooling of the tropical pacific (Trenberth, 1997). During the El Nino Southern Oscillation (ENSO), there tends to be, but is not necessarily strictly, a lower abundance in rainfall along the east coast of South Africa, leading to the possibility of drought. However, there have been periods where the ENSO has resulted in an increase of rainfall. The La Nina Southern Oscillation tends to see an increase in rainfall along the east coast of South Africa.

2.2.2 The West Coast of South Africa

The western coast of South Africa is much drier due to the cold Benguela current that flows toward the equator. Low levels of evaporation lead to little precipitation along the coast. Due to dry conditions and little rainfall, the formation of deserts along the west coast of South Africa has occurred. Toward the northwest, the Namaqualand area can be found which runs down the coast from the border of South Africa toward the Olifants river near Eland's Bay. This area is strongly influenced by cold upwelling that occurs off the west coast of South Africa which results in fairly mild desert conditions. A unique feature of the Namaqualand is the reliability of the winter rainfall (Cowling et. al., 1998).

2.2.3 The Interior of South Africa

Within the boundaries of South Africa, changes in rainfall seasonality are also evident. Trends throughout the country on an annual basis appear to not be consistent, however, there are spatial areas which do show significant trends (Kruger, 2006). Six particular regions over the country, including the south coast of the Western Cape, have been identified as showing a significant decrease in annual rainfall. Two regions have shown an increase in annual precipitation levels. du Plessis et. al. (2017) also analyzed rainfall data on a more local scale, focusing on only the Western Cape. The Western Cape was divided into three regions, the Mediterranean, the Karoo and the South Coast. Again, it was difficult to find any statistically significant trends, however, the South Coast region pointed toward a later ending in the rainfall season whilst the other two pointed toward an earlier ending of the season indicating some seasonality change.

2.3 Methods for seasonal change detection

Many different methods have been used in order to develop models with high rainfall prediction accuracy as well as to detect seasonal changes. In order to detect seasonal changes, simple methods as well as more

complex methods can be used. Typically, current data is compared to older data using various methods in order to investigate whether there has been any change. du Plessis et. al. (2017) made use of time lag plots, cumulative plot analysis and moving average analysis in order to detect seasonal change in the Western Cape. Pohl et. al. (2017) made use of the comparison of means for two samples to try detect any significant changes. Kruger (2006) analysed linear trends of a variety of precipitation indices. Dourte et. al. (2015) investigated rainfall trends using fixed threshold and percentile-defined threshold. Various modelling methods have also been used. Lumsden et. al. (2009) made use of General Circulation Models to investigate any changes in the hydrological cycle.

Percentage mean cumulative rainfall has also been used for both the rainfall amount as well as the changes in the number of rainy days. Odenkule (2004) investigated onset and cessation dates in Nigeria using these methods. More recently, Abrahams (2019), made use of a hybrid method incorporating the percentage cumulative distributions used by Odenkule (2004) as well as exceedance thresholds.

Hidden Markov Models (HMMs) have been used to model rainfall. HMMs are an extension of Markov models in the case where the observation is a probabilistic function of the state. Although typically used for speech recognition and signal-processing, HMMs are relatively versatile and can be used to model day to day rainfall predictions (Yaman, 2014).

2.4 Temperature Forecasting

Many studies have been performed analyzing other climate variables globally and over southern Africa. Owing to the increased awareness of 'global warming' and 'climate change', temperature is one such variable that has had more research into it. According to Hansen (2006), global temperature is a popular tool to measure the stability of the global climate. The change in temperature is also transforming at a much more constant rate and is thus much easier to forecast (Barnett, 2005).

The global temperature has been increasing by approximately 0.13 degrees Centigrade per decade since 1950 and is expected to increase at 0.2 degrees Centigrade per decade for the next few decades (Lobell et. al., 2011). Kruger et. al. (2006) analyzed data for various locations within South Africa where data quality obtained was high. All stations showed an increase in temperature, with a bigger increase having been seen in the interior of the country. As temperatures globally continue to increase, more and more extreme weather events will be experienced. Rainfall is likely to become increasingly intense in its falling patterns.

2.5 Environmental Impacts - catchment dams

Many environmental impacts are being experienced as a result of a change in rainfall patterns. Rainfall has a major impact on crop health as well as on the underlying soil structure of the land. The impacts on agriculture could have disastrous affects as rainfall and temperature are key for agriculture production. Food prices in many areas are already seeing rapid increases as a result of increasing drought periods resulting in more difficult crop production conditions (Lobell et. al., 2011). With a population that is continuing to grow and crop production becoming more and more important, this is a worrying factor to consider.

Barnett (2005) highlights that more than half of the world's potable water supply comes from rivers - whether that be collected directly from the river or a reservoir. Both precipitation and temperature changes have significant effects on river runoff. A decrease in precipitation typically results in a decrease in volume of runoff whilst temperature changes tend to affect the timing of runoff.

Severe soil erosion is another potential impact from a changing rainfall pattern. Severe soil erosion can lead to the obstruction of sustainable management of soil and water resources (Zhang et. al., 2005). Soil erosion will also hamper any long-term planning of soil and water conservation in an area. With an increase in the number of intense rainfall events being experienced, an increase in soil erosion is likely to occur. Analysis on soil erosion has been performed for the Yellow River basin in China. An expected increase in soil erosion due to increasing rainfall events has resulted in much more attention been given to water and soil conservation practices - vegetation rehabilitation and check-dam construction (Zhang et. al., 2005).

2.7 Summary

Many impacts have been experienced globally due to changes in rainfall and climate patterns and it is important to study these. Many different methods have been used when attempting to analyse seasonality over time. Chapter 2 highlighted these methods along with the climate structure of the Western Cape. Chapter 3 will highlight the methodology used to cluster the weather stations, determine start and end season days as well as the methods used to analyse these days.

3. Methodology

The aim of this chapter is to investigate changes in seasonality in rainfall for thirty weather stations from the Western Cape. This chapter lays out the different methods that have been used in order to detect changes in seasonality. Initially, the weather stations are clustered based on their annual total rainfall amounts and monthly average rainfall amounts. Clustering is important as analysis is performed per cluster, rather than by weather station. Next, a threshold approach, set out criteria and Generalized Additive Models are used to estimate the start and end of the rainfall season for each cluster.

3.1 The Data

Data was acquired from the South African Weather Service (SAWS) from 30 weather stations within the Western Cape region. The weather stations are spatially chosen from a possible 295 active stations in order to obtain a general overview for the whole of the Western Cape. All of the weather stations have daily rainfall data with most of the data starting on 1 January 1918 and ending 31 December 2017.

Table 1: 30 weather stations for which daily rainfall data was obtained from South African Weather Service.

Weather Station	Latitude	Longitude	Years Missing
Beaufort Wes Stolshoek	-32.33	22.49	1918 to 1928,1973 to 1978
Cape Agulhas	-34.83	20.01	
Cape Columbine	-32.83	17.86	1918 to 1935
Cape Point	-34.35	18.49	
Coetzeeskraal	-32.00	24.02	1918 to 1931
De Doorns	-33.47	19.67	1936 to 1963
Gannakraal	-31.91	22.84	1918 to 1925, 1964 to 1971
Gansbaai Danger Point	-34.63	19.30	1950
Graafwater	-32.15	18.61	1918 to 1923, 1952 to 1956, 1975 to 1978
Grabouw	-34.15	19.02	1956, 1966
Hopefield	-33.07	18.35	
Jonkersberg - Bos	-33.93	22.23	1918
Knysna	-34.04	23.05	1935 to 1950
Kruisrivier	-33.43	21.86	1918 to 1927
Ladismith	-33.50	21.27	
Langebaan	-33.09	18.03	
Langgewens	-33.28	18.71	1918 to 1929
Merweville - Pol	-32.66	21.52	
Molteno	-33.94	18.41	1925 to 1954
Montagu	-33.78	20.13	1949, 1952 to 1955, 1957, 1958, 1960, 1967, 1968
Nuwerus	-31.15	18.36	1918 to 1924
Paarl	-33.72	18.97	1987 to 1992
Piketberg-SAPD	-32.91	18.75	
Prince Albert	-33.22	22.03	
Quaggasdrift	-32.10	24.02	1918 to 1931
Rhebokskraal	-33.99	19.83	1918 to 1928
Robben Island	-33.81	18.37	
Rondawel	-33.20	22.66	
S A Astronomical Observatory	-33.93	18.48	
Vanrhynsdorp	-31.61	18.75	1919 to 1921

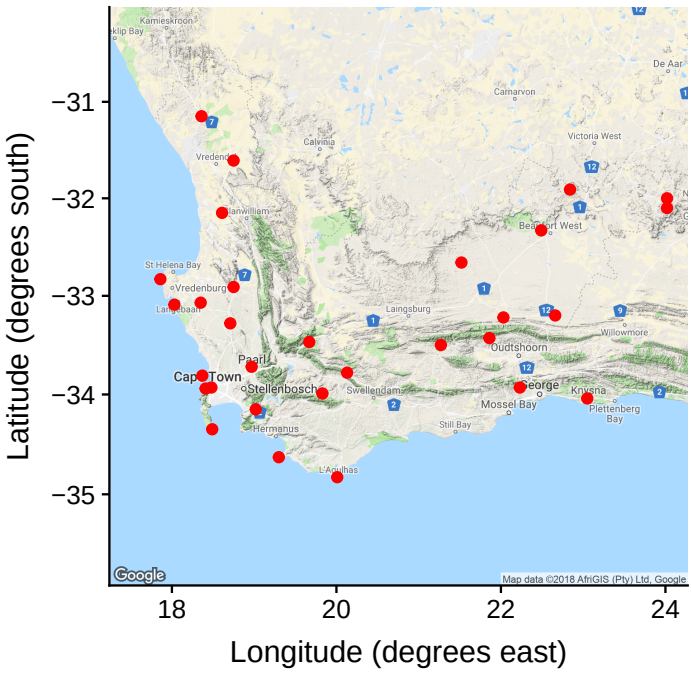


Figure 2: Map of the Western Cape with chosen weather stations indicated by red dots. Source: Google Maps

3.2 Classifying the Weather Stations

3.2.1 Cluster Analysis

Due to the spatial distribution of the weather stations and the underlying climate structure of the Western Cape, many of the weather stations are likely to have different seasonal rainfall patterns. As a result of these different seasonal patterns, it is unreasonable to apply the exact same methodology to all of the weather stations and so varying analysis is performed based on individual patterns of weather stations. In order to account for these differences in rainfall seasonal patterns and to group similar weather stations together, cluster analysis is performed. Dealing with clusters as opposed to individual weather stations also helps remove some of the variation as it naturally smooths the data (Alam et. al., 2019). Both hierarchical and k-means clustering algorithms are used.

Based on knowledge of the Western Cape, there are at least three different regions which experience different rainfall seasonality, these being the 'Mediterranean Region', the 'South Coast Region' and 'The Karoo' (du Plessis and Schloms, 2017). Therefore, at least three clusters should be determined.

The purpose of the clustering is to group weather stations together that have similar total annual rainfall trends as well as similar seasonal trends within each year. Some of the weather stations have low summer rainfall and high winter rainfall, some have high summer rainfall and low winter rainfall whilst others are aseasonal, having fairly constant rainfall throughout the year.

If clustering using only annual total rainfall amounts, weather stations with similar annual total rainfall trends over the years will be grouped together. This clustering approach accounts for years of drought and flooding, however, it does not account for seasonal trends within the year and so weather stations with very different seasonal trends could be grouped together if their total annual rainfall amounts are similar. If clustering using the average monthly rainfall, seasonal trends within the year are accounted for, however, the trends over time are negated. The average monthly rainfall is computed by aggregating the rainfall for each month for each year. These values are then averaged over the total number of years.

In order to perform the clustering, a matrix is setup containing data which the clusters are formed on. Two different matrices are used to perform clustering. The first matrix has each weather station as a row of the matrix and then the annual total rainfall for each station as the columns - this results in a 30 by 100 matrix (the '30' represents the 30 weather stations and the '100' represents the 100 years of the data). The second matrix has all the weather stations as the rows, but now the columns are the average monthly rainfall amounts for each weather station - this gives a 30 by 12 matrix (the '12' representing the 12 months of the year). All the values in these matrices are absolute values.

Cluster analysis requires a complete data set with no missing values. As there are missing values from many of the weather stations in the case where the total annual rainfall is being used to cluster on, two methods are used to deal with the missing values. Firstly, if a column has one or more 'NA' values, the whole column is completely removed. However, once these values have been removed, there are only 35 years remaining of data of the original 100 years. This is far from ideal as lots of potential information is left out. Thus, multiple imputation is used. The purpose of multiple imputation is to, as accurately as possible, impute values where data values are missing. This will result in a complete data set with 100 years worth of data for all the 30 weather stations.

3.2.2 Imputing the Missing Values

Multiple Imputation (MI) is used to calculate total annual rainfall amounts for the years where there are missing values. This is necessary as both the clustering algorithms require a complete data set. These missing values are only used for performing the clustering - they are not used in the analysis of the start and end of season. Multiple imputation works by creating 'complete' multiple data sets where the missing values are predicted multiple times. The predicted values are imputed based on the observed data for the variable (weather station) as well as the observed data from the other variables (all the other weather stations) (Schafer et. al., 2002). As the multiple predictions are aggregated for each missing value, uncertainty for the predictions are taken into account (Azur et. al., 2012).

The total rainfall amounts from each year from all the weather stations are split up into a matrix where the rows correspond to each weather station and the columns correspond to all the years in the data (1918 to 2017). Thus, a missing value is imputed by looking through the annual total rainfall amounts of all the weather stations in that same year (column of the matrix) as well as the annual total rainfall amounts of the weather station for all years (row of the matrix). So, based on the annual total rainfall amounts of all the weather stations for a specific year as well as the annual total rainfall amounts for all the years for that specific weather station, the missing value is imputed.

This method is implemented using the 'Mice' (Multiple Imputation Chained Equations) package (van Buuren et. al., 2011) in 'R' using the predictive mean matching (PMM) method. 'Mice' assumes that the data are missing at random (MAR). The imputation follows a chained equation approach where each variable (weather station) is modeled according to its distribution conditional upon the other variables in the data. Initially, starting estimates are obtained for the missing values for each variable by using the mean of each variable. Each variable is then modeled conditionally on the other variables and the missing values are predicted from the model and then updated and replaced. This process is then repeated. Each iteration forms new models based on the updated predicted missing values until convergence or a specified number of iterations occurs. The observed values remain the same throughout the whole process - only the missing values are updated after each iteration. Once the last cycle is performed, the final predicted missing values are used (Azur et. al., 2012).

Predictive Mean Matching (PMM) with five iterations is used when performing the MI. PMM is similar to the regression method and it is a semi-parametric imputation approach. What makes it different to the regression method is that for each missing value, "it fills in a value randomly from among the observed donor values from an observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model" (Heitjan et. al., 1991; Schenker et. al., 1996).

3.2.3 Hierarchical Clustering

Hierarchical clustering is the first method used to compute and generate clusters. The advantage of hierarchical clustering is that the number of clusters does not need to be specified prior to performing the clustering. Agglomerative (bottom up) clustering is used to determine the clusters. Firstly, a 'dissimilarity matrix' is computed based on the Euclidean distances between weather stations in terms of rainfall. Agglomerative clustering works by initially viewing each weather station as a cluster. Two clusters are then combined based on their Euclidean distances (the more similar two clusters are, the more likely that they are placed together).

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}, \tag{1}$$

where $\mathbf{p}$ and $\mathbf{q}$ are two weather stations and the i^s represent either the year or the month depending on which data is used.

This process is repeated until there is only one cluster which contains all of the weather stations (Johnson et. al., 2007). The weather stations are treated as the cases (rows of the matrix) whilst either the annual total rainfall or the average monthly rainfall are the columns of the matrix. This algorithm is implemented in 'R' using the 'cluster' package (Maechler et. al., 2018).

Clustering is performed on the *imputed annual total rainfall* data as well as the *monthly average rainfall* data. Once clustering is completed, a dendogram plot is produced. The dendogram plot shows the dissimilarity of all the formed clusters, starting from every weather station being its own cluster to there being only one cluster including all the weather stations. By looking at this plot, the optimal number of clusters that should be chosen is obtained. The position where there is a big difference in dissimilarity between number of clusters is chosen as the cut off point. The number of branches at this point then equates to the number of optimal clusters.

Once the optimal number of clusters is determined, the data from the weather stations in each cluster is grouped together. A plot is then produced showing the average monthly rainfall averaged over all years of each weather station as well as the combined average monthly rainfall of all the weather stations in that specific cluster. If there is a weather station with an average monthly rainfall trend that is very different to the average monthly rainfall of the cluster, more clustering is performed/another cluster is added.

3.2.4 K-means Clustering

Next, k-means clustering is performed. Clustering is first performed on the *imputed annual total rainfall* data of each weather station and then on the *monthly average rainfall* data of each weather station. The aim of k-means clustering is to minimize the intra-cluster variance:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^j - c_j||^2, \tag{2}$$

where $\mathbf{c}$ represents the centroid values.

K-means clustering follows a specific algorithm in order for the clusters to be computed. Firstly, the number of clusters is chosen prior to any analysis as well as initial centroid values of these clusters. When doing this process in 'R', the centroid values are randomly generated and then the clustering is performed. The points (weather stations) are then assigned to a cluster based on the shortest Euclidean distance between the centroids of the predefined clusters and the points.

$$d(\mathbf{x}, \mathbf{c}) = \sqrt{\sum_{i=1}^{n} (x_i - c_j)^2}, \tag{3}$$

where $\mathbf{c}$ represents the centroid values.

Once all these points have been assigned to a cluster, the centroids of the clusters are recalculated based on the mean of all the points in the cluster.

$$d(\mathbf{c}^{(\mu+1)}) = \frac{1}{|c_i^{(\mu)}|} \sum \mathbf{x}, \tag{4}$$

where $|c_i^{(\mu)}|$ is the number of vectors assigned to cluster $c_i^{(\mu)}$.

The points are once again assigned to a cluster based on the shortest Euclidean distance to the new centroids. This process is repeated until convergence is achieved (there is no change in the centroid values) or until a specified stopping point (Johnson et. al., 2007). This algorithm is implemented in 'R' using the "cluster" package.

In order to determine the optimal number of clusters, the "Elbow Method" for k-Means clustering is used. This method plots the *total within-clusters sum of squares* for the minimum (every weather station is in one cluster) to the maximum (every weather station is its own cluster) number of clusters. As this is a visual method, the optimal number of clusters is selected based on a point on the plot where the rate of change of the *total within-clusters sum of squares* starts to flatten out.

As for hierarchical clustering, once the clusters are determined, they are grouped together and average monthly rainfall plots are produced for each weather station in the cluster as well as the average monthly rainfall of the whole cluster. These plots are used to visually identify whether there are any weather stations that appear to be highly dissimilar from the cluster's average pattern. If there are discrepancies between the stations within the cluster, another cluster may be added to see if this results in more uniform clusters.

3.2.5 Computing the *Cluster Daily Average Rainfall*

Once the most visually optimal clusters are chosen, new daily rainfall amounts are calculated for each cluster. The daily rainfall amounts from each weather station for every specific day within the one hundred years are added together for all the weather stations within the cluster. These values are now averaged giving a new *cluster daily average rainfall* amount for each specific day of the hundred year time period. This is done for every day for every year within the data set. This averaging process helps in reducing the affect of potential outlying events and reducing some of the variation in the data set whilst still including high levels of information of the data set (Alam et. al., 2019). This averaging process also helps deal with missing values in the data set. These new *cluster daily average rainfall* amounts are the data that is used for all of the analysis.

3.3 Defining the Different Terms used for the Different Data Sets

Table 2: Definitions of different data sets used.

cluster daily average rainfall	Refers to the daily rainfall amount from clusters. This value is calculated by averaging the daily rainfall amount for each specific day in the time period over all weather stations in each cluster.
annual total rainfall	Refers to the total rainfall for each year within the time period.
monthly average rainfall	This value is calculated by summing the daily rainfall amounts for each month for every year for each weather station. Next, for every month of the year, the average monthly rainfall is calculated by averaging the monthly rainfall over all years.
cluster monthly average rainfall	This value is calculated by taking the *monthly average rainfall* and averaging over all weather station within the cluster.
imputed annual total rainfall	Refers to the total annual rainfall for each year within the time period. Years that are missing have been imputed using multiple imputation so that there is a value for all of the 100 years.

3.4 Cumulative Plots

Cumulative plots are produced in order to obtain a visual representation of rainfall amounts for different years for each cluster. Cumulative plots are a good way to detect differences between distributions or shifts in time of distributions. For the first cumulative plots, three years are looked at that are approximately 50 years apart (1918, 1968 and 2017). These years represent the first, approximately the middle and the last year of the data set. These plots are produced using the *cluster daily average rainfall* data (*Table 2*).

For the next cumulative plots, the first ten years of data (1918 to 1927) is averaged for each specific day. This is done by taking the *cluster daily average rainfall* data for the cluster for the first ten years and then averaging over each specific day in the year giving 365 values. These values are then smoothed by fitting a cubic smoothing spline to the data. These values are then cumulated in order to obtain one cumulative curve. The same is done for the last ten years of the data set (2008 to 2017), and lastly, an average for the whole data set is computed (1918 to 2017). These three cumulative curves are then plotted and compared to one another. From these plots it is seen if rainfall amounts are changing (amount of separation of the lines at the end of the curve) as well as whether rainfall appears to be raining in different time periods within the year (shape of the curves).

3.5 Seasonal Plots

Seasonal plots are produced to give a visual representation of seasonal patterns over time for different clusters. Similar to the cumulative plots, firstly three chosen years are plotted (1918, 1968 and 2017). These years represent the first, approximately the middle and the last year of the data set. These lines are based on the *cluster daily average rainfall* data.

Next, the first ten years (1918 to 1927) and last ten years (2008 to 2017) as well as the whole *cluster daily average rainfall* data set (1918 to 2017) are aggregated and averaged for each day in the year per cluster.

These average rainfall amounts are then smoothed by fitting a cubic smoothing spline to the data. These values are then plotted to give an average seasonal profile of the first ten, last ten years and average of the data. These curves provide a way to visually illustrate shifts and changes in seasonality, such as shorter seasons as well as changes or shifts in the peak of rainfall.

3.6 Start and End of the Rainfall Season

The start and end of the rainy season is estimated for each cluster using a variety of methods. The first approach used makes use of thresholds, next, predefined criteria are determined and lastly, Generalized Additive Models are used to estimate the start and end of the rainfall season.

3.6.1 A Theshold Value Based on Cumulative Annual Total Rainfall

Firstly, a threshold is decided for each cluster based on the methods implemented by Abrahams (2019) when trying to estimate the onset/cessation of the rainfall season for weather stations within the Western Cape. A different threshold is found for each cluster as each cluster has its own unique rainfall amounts and trends over time. Once this threshold is found, the start/end of the rainfall season is estimated.

In order to choose the threshold for each cluster, firstly, the daily rainfall amounts over all years (1918 to 2017) and weather stations within each cluster are averaged. This results in an average daily rainfall (averaged over all years and all weather stations in the cluster). Next, the cumulative anomalies are found - the anomalies are the difference between the daily mean rainfall and the mean of these values. These anomaly values are then cumulated and the days of the year on which the minimum and maximum cumulative anomalies occur are found. Cumulative plots are then produced of the daily mean rainfall and the threshold rainfall amount that will define the start/end of season is found by finding the cumulative rainfall values corresponding to the days found using the cumulative anomalies (Abrahams, 2019).

For example, the start of season could be defined once 100 mm of the total cumulative rainfall for the year has fallen and the end of the season can be found once 400 mm of the total cumulative rainfall for the year has fallen. This is then done for every year for each cluster and comparisons are then made to see whether there have been any changes in the start/end of the rainfall season.

Due to the variability of daily rainfall, this is possibly a naive way to estimate the start and end of the season as the estimates are strongly influenced by single events. A single outlying early/late rainfall event can cause the start/end of season to shift by days resulting in highly variable estimates of start of season. Particularly in regions where there is low annual rainfall amount, outlying events can easily cause the threshold to be crossed much earlier than the actual start of the rainfall season or much later than the actual end of season. These outlying events would then also skew the estimated length of the rainfall season. A smoothing method that puts less weight on single events may be preferable.

3.6.2 Using Criteria

Next, a more rigorous process is used to compute the start and end of seasons. Instead of a threshold value being crossed, a few criteria need to be satisfied in order for the rainfall season to be estimated to have started or ended. As the weather stations are clustered, the criteria will be adjusted based on the rainfall patterns of each specific cluster. An example of criteria for a cluster that has a winter rainfall season is listed below. These criteria are based on research done by Laux et. al. (2008) for predicting the beginning of the rainfall season in West Africa. This definition was initially set out by Stern et. al. (1981).

Initially, Stern's criteria are used on the clusters that experience winter rainfall in order to estimate the onset date of the rainfall season. Stern came up with the following criteria for the winter rainfall region he was studying in West Africa:

- 25 mm of rain needs to have fallen within a 5 day period.

- Within that 5 day period, rain needs to be captured on at least 3 of those days.
- Within the next 30 days following this 5 day period, there can be no period of 7 consecutive days or longer where no rain is captured.

If all criteria are met, the first day of the initial five day period is estimated as the start/end of season.

Modified criteria have been used in the past. Sarria-Dodd and Jolliffe (2001) devised their own modified criteria set out by Stern et. al. (1981) when attempting to predict the onset date of rainfall in Burkina Faso. Sarria-Dodd and Jolliffe (2001) decided to modify the definition by relaxing certain criteria as onset dates were being supplied that seemed too late to be reasonable. By relaxing certain criteria, the onset day will likely occur earlier in the year. To the same affect, definitions are modified for each cluster based on knowledge of the area and the typical rainfall conditions.

As for many years the criteria set out by Stern et. al. (1981) are not met for clusters 2,4 and 5, the criteria are modified for each cluster based on their rainfall trends and typical amounts for the year. This process should be more accurate in estimating the start of the season as it is not influenced by outlying events. Criteria are modified based on Stern's three criteria with the addition of one new criteria. This new criterion is that a certain percentage of the cumulative rainfall needs to have fallen before the rainfall season is estimated as well as a certain percentage can not have fallen before the end of the rainfall season is estimated. For the first criterion set out by Stern et. al. (1981), values are chosen based on the plotted cumulative curves.

Winter Rainfall Clusters:

Criteria for Start of Season:

Cluster 1 (South Mediterranean Region):

- 20 mm of rain needs to fall within a 5 day period.
- within that 5 day period, rain needs to be captured on at least 3 of those days.
- within the next 40 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 15% and less than 50% of the total rainfall for the year must have fallen.

Cluster 3 (Central Mediterranean Region):

- 25 mm of rain needs to fall within a 5 day period.
- within that 5 day period, rain needs to be captured on at least 3 of those days.
- within the next 30 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 15% and less than 50% of the total rainfall for the year must have fallen.

Cluster 4 (West Mediterranean Region):

- 10 mm of rain needs to have fallen within a 6 day period.
- within that 6 day period, rain needs to be captured on at least 3 of those days.
- within the next 30 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 15% and less than 50% of the total rainfall for the year must have fallen.

Criteria for End of Season:

Cluster 1 (South Mediterranean Region):

- 12 mm of rains need to fall within a 6 day period.
- within that 6 day period, rain needs to be captured on at least 3 days.

- within the next 30 days following this 6 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 75% of the total rainfall for the year must have fallen.

Cluster 3 (Central Mediterranean Region):

- 20 mm of rain needs to fall within a 5 day period.
- within that 5 day period, rain needs to be captured on at least 3 days.
- within the next 23 days following this 5 day period, there can be no period of 7 consecutive days or longer where no rain is captured.
- more than 75% of the total rainfall for the year must have fallen.

Cluster 4 (West Mediterranean Region):

- 7.5 mm of rain needs to fall within a 6 day period.
- within that 5 day period, rain needs to be captured on at least 3 days.
- within the next 30 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 75% of the total rainfall for the year must have fallen.

Summer Rainfall/Aseasonal Clusters:

Criteria for Start of Season:

Cluster 2 (South Coast Region):

- 20 mm of rain needs to fall within a 5 day period.
- within that 5 day period, rain needs to be captured on at least 3 days.
- within the next 30 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 15% and less than 50% of the total rainfall for the year must have fallen.

Cluster 5 (Karoo Region):

- 6 mm of rain needs to fall within a 6 day period.
- within that 6 day period, rain needs to be captured on at least 3 days.
- within the next 25 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 15% and less than 50% of the total rainfall for the year must have fallen.

Criteria for End of Season:

Cluster 2 (South Coast Region):

- 12 mm of rain needs to fall within a 5 day period.
- within that 6 day period, rain needs to be captured on at least 3 days.
- within the next 30 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 75% of the total rainfall for the year must have fallen.

Cluster 5 (Karoo Region):

- 7.5 mm of rain needs to fall within a 6 day period.
- within that 5 day period, rain needs to be captured on at least 3 days.
- within the next 30 days following this 5 day period, there can be no period of 10 consecutive days or longer where no rain is captured.
- more than 75% of the total rainfall for the year must have fallen.

3.6.3 Generalized Additive Models

As a third method, generalized additive models (GAMs) are used to estimate start and end of season dates.

GAMs are used because they model the relationship between parameters (such as mean and probability of zero rainfall) and predictor variables, such as time of year. It is reasonable to assume that these parameters change smoothly over time, and that the observations are a noisy realization of this smooth underlying process. Even though rain does not fall as a smooth function throughout the year, it is assumed that there is an underlying unobserved mean daily rainfall amount that does and that noisy data points are then observed away from this mean. This smooth function for the mean represents the conditional mean - given that it does rain. By estimating this mean daily rainfall function, it can then be used to determine the start and end of season.

One can also define a GAM with a parameter for the probability of rainfall. This is because the Zero Adjusted Gamma (ZAGA) model is used, assuming a Gamma distribution for the rainfall and an additional component to model days without rainfall. It is also assumed that there is an underlying unobserved probability of zero rainfall for each day of the year. This parameter is also a smooth function with noisy data points being observed away from this function. The only predictor variable that is used, is time (*days of the year*). Thus, the smooth functions that are obtained explain how the mean rainfall and probability of zero rainfall change over time.

A GAM is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates (Hastie and Tibshirani, 1990). Just like any generalized model, some transformation of the parameters links the parameter with the explanatory/predictor variables giving the following model:

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_m(x_m), \tag{5}$$

where $f_j(x_j)$ are smooth functions and Y_i has an exponential family distribution. $Y_i \sim Exponential\ Family(\mu_i, \phi)$ with mean μ_i and scale ϕ (Wood, 2017).

Basis expansions are used to estimate these smooth functions ($f_j(x_j)$) of the model. A basis is made up of basis functions which are known.

$$f(x) = \sum_{k=1}^{K} b_k(x)\beta_k, \tag{6}$$

where $b_j(x)$ is the j^{th} basis function and β_j is the unknown parameter. If using linear regression, equation (6) can be substituted into a simple linear model.

$$y_i = f(x_i) + \epsilon_i \tag{7}$$

For example, if f were a 4th order polynomial, the basis space would be as follows: $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$, $b_4(x) = x^3$ and $b_5(x) = x^4$. Equation (6) would then become

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5, \tag{8}$$

and the following simple linear model is formed when substituting into equation (7):

$$y_i = \beta_1 + \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5 + \epsilon_i \tag{9}$$

However, polynomial bases are not always the best at estimating unknown functions. This is a result of Taylor's theorem which implies that a polynomial basis is useful for cases where a single point is the focus within the vicinity of f (Wood, 2017). Thus, splines will be used. GAMs can be formed using cubic spline bases. Any smooth function can be used as a basis, but the cubic spline is often used.

Splines are produced by fitting a specified number of piecewise polynomials together. The number of polynomials is determined by the number of knots chosen. The knot is the point where two polynomials are joined. If there are n knots, there will be n+1 polynomials. Three different parameters can be modeled when using the ZAGA distribution - the 'variance', the 'mean' as well as the 'probability of zero'. However, for the purpose of this thesis, only the 'mean' and 'probability of zero' are chosen to be modeled. The 'mean' refers to the expected rainfall to occur on each specific day of the year given rain is observed and the 'probability of zero' is the probability that zero rainfall is captured on each specific day of the year.

Because rainfall is seasonal, a periodic cubic spline basis is used. This ensures that the modeled mean daily rainfall starts and ends on the same value. Periodic splines make use of B-spline basis functions to define piece-wise polynomials. In order to produce a smooth model and to avoid over fitting, the coefficients of the basis functions are penalized (Rigby et. al., 2007). The smoothness of the model is controlled by penalizing the wiggliness. Instead of fitting the model by minimizing

$$||y - X\beta||^2, \tag{10}$$

(which is for a linear regression model), the model can instead be fit by minimizing

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx \tag{11}$$

(Erni, 2019).

Minimizing this penalized sum of squares forces the parameter estimates to be a compromise between fit on the left hand side (minimize error sum of squares) and wiggliness on the right hand side, and will result in smoother estimated splines. The more wiggly f is, the higher the values the penalty will take on in order to make the function smoother. If f is smooth, the penalty will take on much smaller values. λ, the smoothing parameter, controls the trade-off between a smooth function and a function that over fits to the data.

If the distribution of the response is not normal, the GAM is fit using penalized iteratively reweighted least squares (Wood, 2017). The following process is followed to obtain the estimates for the model:

1. Given that the vector μ is the corresponding estimated mean for the current linear predictor estimate vector, $\hat{\eta}$, the following must be calculated:

$$w_i = \frac{1}{V(\hat{\mu}_i)g'(\hat{\mu}_i)^2} and z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i, \tag{12}$$

where $var(Y_i) = V(\mu_i)\phi$, g is the link function, w_i are the weights and z_i are the working residuals.

2. Next, $\mathbf{W}$ is defined as the diagonal matrix such that $W_{ii} = w_i$. Minimize

$$||\sqrt{\mathbf{W}}\mathbf{z} - \sqrt{\mathbf{W}}\mathbf{X}\beta||^2 + \lambda_1\beta^T S_1\beta + \lambda_2\beta^T S_2\beta \tag{13}$$

with respect to β resulting in a new estimate for $\hat{\beta}$ and therefore, new updated estimates for $\hat{\eta} = \mathbf{X}\hat{\beta}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$

This process is repeated until convergence.

The 'gamlss' package (Rigby and Stasinopoulos, 2005) is used in 'R' in order to fit GAMs. This package allows for a variety of underlying distributions to be used. As the daily rainfall data is filled with zeros and gamma distributed (*Figure 4*), the Zero Adjusted Gamma Distribution (ZAGA) is used when performing the modelling. The ZAGA model is different to the Gamma model as it allows for extra zeros in the observations.

This allows the probability of zero rainfall to be modeled. Similar to the assumption of there being a smooth mean daily rainfall function, it is assumed that there is a smooth probability of zero rainfall function. This smooth function is also used to estimate the start and end of the rainfall season.

The probability distribution function of the ZAGA is defined as follows:

$$f_Y(y|\mu,\sigma,\pi) = \begin{cases} \pi & \text{if } y = 0 \\ (1-\pi)\left[\frac{1}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}} \frac{y^{\frac{1}{\sigma^2}-1} e^{-\frac{y}{(\sigma^2\mu)}}}{\Gamma(\frac{1}{\sigma^2})}\right] & \text{if } y > 0 \end{cases} \tag{14}$$

for $0 \leq y < \infty$, $0 < \pi < 1$, $\mu > 0$, $\sigma > 0$,

where μ denotes the conditional mean (expected amount of rainfall if it rains), σ the variance and π the probability of zero. $E(Y) = (1-\pi)\mu$ and $Var(Y) = (1-\pi)\mu^2(\pi + \sigma^2)$.

In addition,

$log(\mu) = f_1(day)$ $logit(p_i) = f_2(day)$

are used to model changing mean rainfall and probability of rainfall over the season.

The above ZAGA model is used to model the conditional mean daily rainfall amount as well as the probability of zero rainfall for each cluster for every year. A smooth model is required for each year. By varying the degree of knots, the smoothness of the splines can be controlled. For these models, six knots are chosen based on a visual inspection of the model. This number of knots is chosen as an increase of knots leads to over fitting of the model and the model becomes too sensitive to outlying events.

A variety of parameter estimates are calculated to determine the model fitted to the cluster average daily rainfall for each year. The parameters that need to be estimated are the spline coefficients for the mean and the probability curve over the course of one year. These parameters are estimated using penalized iteratively reweighted least squares. Using the parameter estimates, two smooth curves are obtained - one describing how the mean amount of rainfall changes over the year, and the other describing how the probability of no rainfall changes over the year. These estimated curves are used in a variety of ways to determine seasonality. Firstly, by using thresholds and secondly, by analyzing the gradient of the model.

3.6.3.1 A Threshold Approach

The first approach used to estimate the start and end of the rainfall season, is a threshold approach. A threshold value is used to estimate the start and end of season for both the 'mean' and 'probability of zero' models. For the 'mean', a threshold is chosen where the mean rainfall on a specific day of the year crosses a certain value. For the 'probability of zero', a threshold is chosen where the probability of no rainfall crosses a chosen value. These thresholds change with each cluster.

3.6.3.2 The Rate of Change of the Mean

Secondly, the difference between consecutive points (the gradient) is investigated. The gradient between consecutive points shows the rate at which the modeled mean rainfall is increasing/decreasing. When using the modeled mean daily rainfall, finding the point at which the gradient increases steeply could be used as an estimate for the start in the rainfall season. Finding the point where there is a steep decrease in the gradient could indicate an end to the season. When using the modeled probability of zero rainfall, a steep decrease in the gradient of the probability of zero rainfall indicates a start in the rainfall season whilst a steep increase in the gradient indicates an end to the season. In the cases where there is no obvious steep change in the gradient, using a threshold can be more depended on to produce a reasonable estimate of start or end of season.

By analyzing the turning points from the rate of change, the point where there is a steep increase in mean daily rainfall and where there is a steep decrease in the probability of zero rainfall is found. This point is

obtained by finding the turning point showing the highest probability of being found at that location. This is determined using the 'pastecs' (Grosjean et. al., 2018) package in 'R'. As most years have many turning points, the search for the turning point is limited between certain days of the year depending on the seasonal trends of the cluster. Turning points are obtained and these points are estimated as the start/end of the rainfall season.

As these points are points of inflection, they are not necessarily the most accurate reflection of the actual start and end of season day. When looking at the start of season, the estimated onset days are likely to be slightly lagged and when looking at the end of season, they are likely to be slightly early. This is because the point of inflection doesn't reflect the point where the gradient suddenly starts to increase or where it starts to plateau after a big decrease (which would indicate the start and end of the season).

3.6.4 Modelling Changes in the Start and End of the Rainfall Season

In this section, the yearly start and end of season dates (1918-2017) are further examined for shifts.

Linear Regression

Firstly, a linear model is fitted to the estimated onset/cessation dates of the rainfall season for the 100 years of data. A linear model determines whether there is any linear trend of the rainfall onset/cessation day. The form of the model:

$$Y = X\beta + \eta \tag{15}$$

Year is the only predictor variable.

The slope of the coefficient indicates change in onset/cessation per year. A gradient that shows evidence of change indicates that there has either been an increase or decrease of the start/end of the rainfall season day (depending on the sign of the gradient). A gradient which shows no evidence of being dissimilar to 0 implies that there has been no linear change of the onset/cessation day.

Generalized Additive Modelling (GAM)

Next, a generalized additive model is used to model changes in the start and end of rainfall season days. The benefit of using GAMs compared to linear regression is that a non-linear model is fit to the data. So, instead of only looking for a linear trend (either increasing or decreasing), non-linear changes are looked for where multiple increases or decreases could have occurred within each time series. Once a GAM is fit to each cluster for the start/end of season estimates, a better visual image is shown of how the start and end season days have shifted and changed over the time series.

Again the 'gamlss' package is used in 'R', however, instead of the ZAGA distribution being used, the normal distribution is used. This is appropriate as differences to the mean onset/cessation day are modeled. A cubic spline basis is used instead of a periodic basis.

3.6.5 Analysing the Models

Now that two different models are determined, the slope from the linear regression model and spline curve are used to identify changes in start/end of season. For the linear model, the gradient of the model is analysed to determine whether it shows any evidence of being dissimilar to zero. The p-value along with 95% confidence intervals are used to determine whether the slope of the linear model is different from zero.

For the GAMs, confidence intervals are obtained for the fitted curve. In order to compute the confidence intervals, the standard error for each fitted value is obtained. This value is then multiplied by some critical value and then added to/subtracted from the fitted value to obtain an interval for each point. 95% confidence intervals are computed and the critical value is found corresponding to the 0.025 quantile of the t-distribution with 98 $(n-2)$ degrees of freedom. As standard errors cannot be found for years where there is no estimate

for onset/cessation day, no confidence interval point can be found for these years. In order to fill these gaps, confidence interval values are estimated by interpolating the missing values using the other generated confidence interval values.

Next, the mean onset/cessation day is computed for each cluster and is plotted along with the 95% confidence interval for the GAM. Areas are now found where the confidence interval of the GAM is either completely above or below the mean onset/cessation day. Years where the mean onset/cessation day is within the confidence interval but is within 5 days of the boundary of the confidence interval are also flagged.

3.7 Analysing the Changes in Duration of the Rainfall Season over Time

Once the start and end of the rainfall season days are calculated, the length of the rainfall season is compared over time by simply looking at the difference of the end and start of season days. However, instead of doing this for every method, the start/end of season day estimates for each cluster are first averaged across all the methods for each year. The difference is then found between the average of the end of season and the average of the start of season.

A linear model and generalized additive model are then fit to these length estimates.

3.8 Summary

Chapter 3 summarized the clustering methods that are used to group weather stations together. The methods that have been used to determine the start and end of season are explained. Lastly, the models that have been used to analyse change - linear models and generalized additive models - are highlighted. Chapter 4 will focus on the results of these methods and models, and in particular, whether there have been any noteworthy changes in seasonality.

4. Results

4.1 Exploratory Data Analysis

4.1.1 Visualising the Response (*Daily Rainfall Amount*)

In order to the illustrate the distributions of the rainfall from the weather stations, a time series for the response (*Daily Rainfall Amount (mm)*) for one of the weather stations (*Ladismith*) for all of the years (1918 to 2017) and then for just one of the years (1918) is plotted as an example. These properties hold across most of the other weather stations.

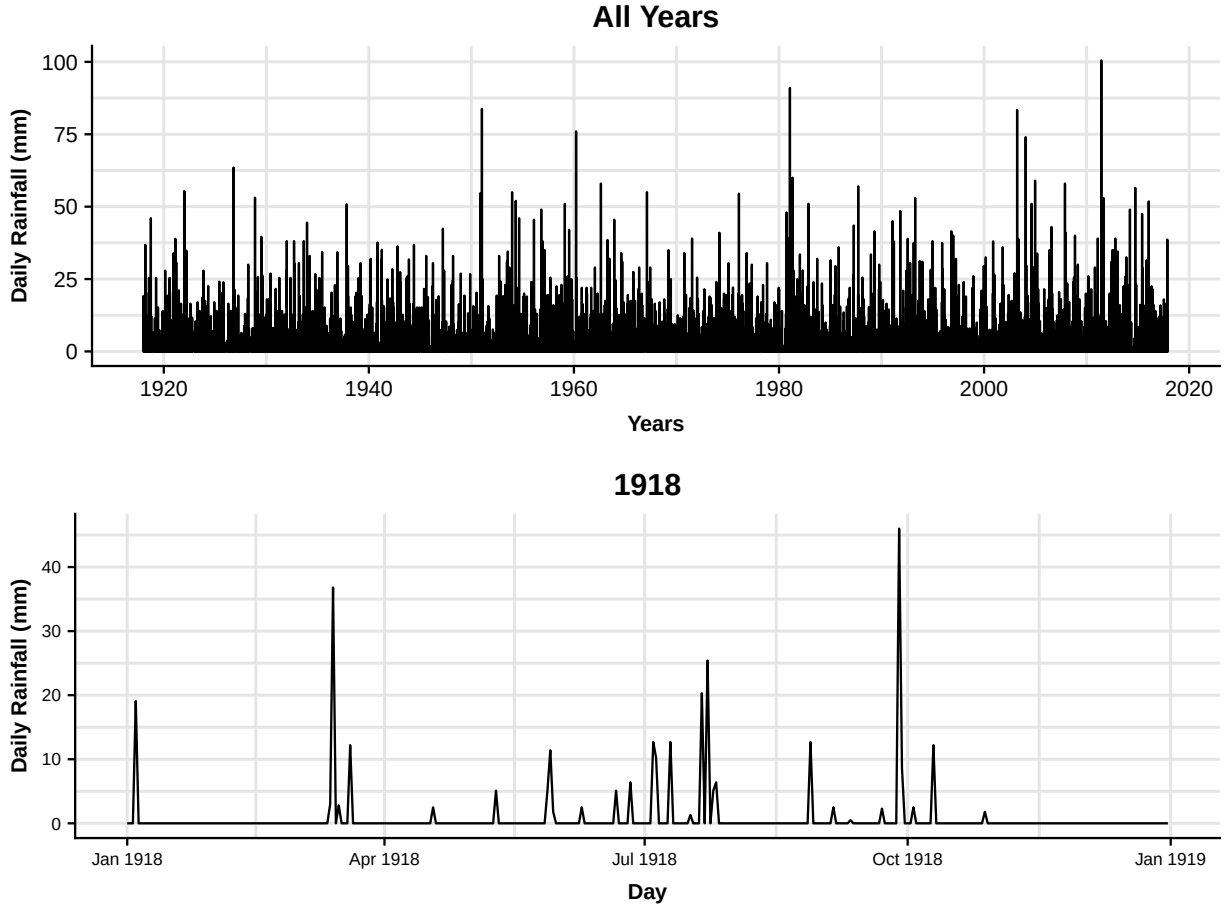


Figure 3: Daily rainfall over time for Ladismith weather station. The top plot represents the years 1918 to 2017 and the bottom plot represents the year 1918.

The long term patterns and maximal events for the *Ladismith* weather station are seen based on *Figure 3 (top)* . The variability of rainfall for a year in this region is seen as well as how the majority of days zero rainfall occurs (*Figure 3, bottom*). Even though this looks like a weather station with a winter rainfall season, this weather station is located in the Karoo of the Western Cape and typically experiences summer/aseasonal rainfall. This plot highlights how variable rainfall is.

Plotting various histograms of the density of the daily rainfall for the *Cape Point* weather station for all of the years (1918 to 2017).

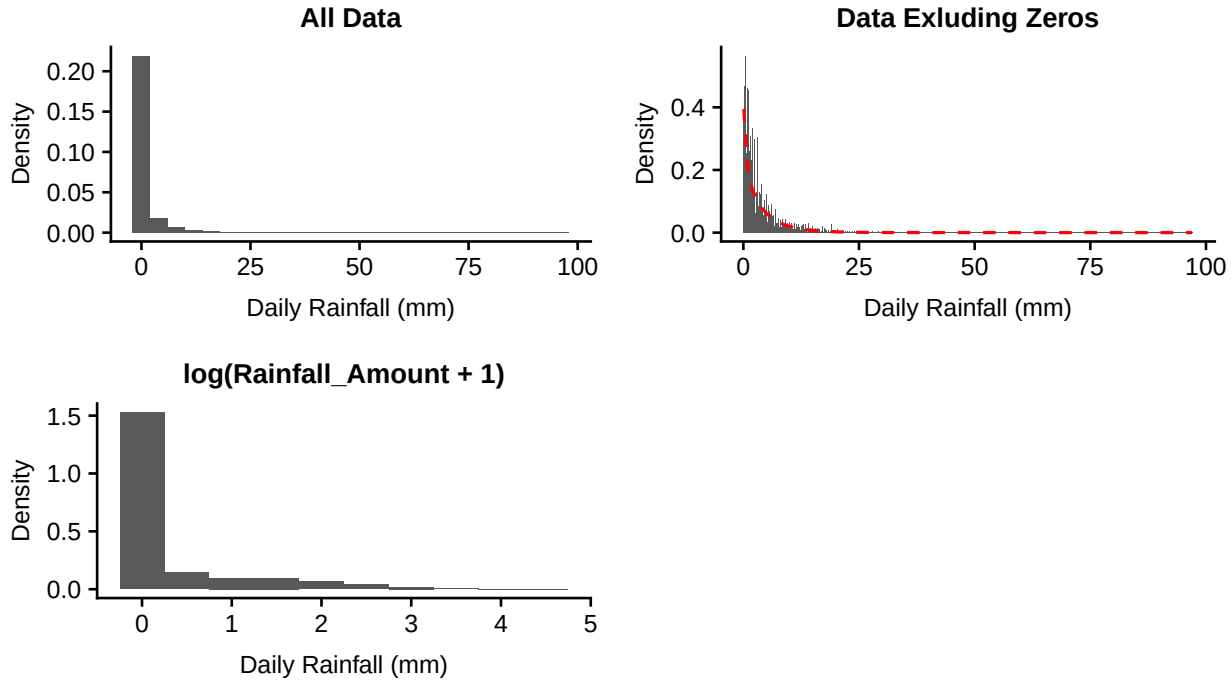


Figure 4: Histograms of daily rainfall amounts using all the data, the data excluding the zeros and the log transform of the data, for the Cape Point weather station for all years (1918 to 2017). The red, dashed line in the top right plot indicates a gamma distribution that has been fit to the data using maximum likelihood.

A very high proportion of the data from this weather station consists of zeros (*Figure 4, top left*) - exactly 26849 out of the 36102 days (74.4%) receive no rainfall. It also is seen that the distribution of the daily rainfall is highly skewed to the right. Even after the the data has been logged transformed (*Figure 4, bottom*), the distribution of the response is still highly skew.

The distribution of the response where all the zeros have been removed is seen based on the top right plot of *Figure 4*. A gamma distribution is fit to the data and it appears to fit the data well. This leads to the possibility of using the zero inflated gamma distribution (ZAGA) when modelling rainfall.

Table 3: Five number summary of daily rainfall amount over all years (1918 to 2017) for the Cape Point weather station.

Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean
0.0	0.0	0.0	0.1	97.0	1.0

Table 3 highlights the highly skewed nature of the data. 75% of the days receive less than 0.1 mm of rainfall. Then, there is a huge jump from the 3rd quartile (0.1 mm) to the maximum (97.0 mm).

Table 4: Five number summary of daily rainfall amount over all years (1918 to 2017) for the Cape Point weather station excluding days of zero rainfall.

Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean
0.1	0.7	2.0	5.0	97.0	4.0

Even with the zeros removed, the data is still highly skewed (*Table 4*).

4.1.2 Visualising the Response (*Daily Rainfall Amount*) for a Cluster

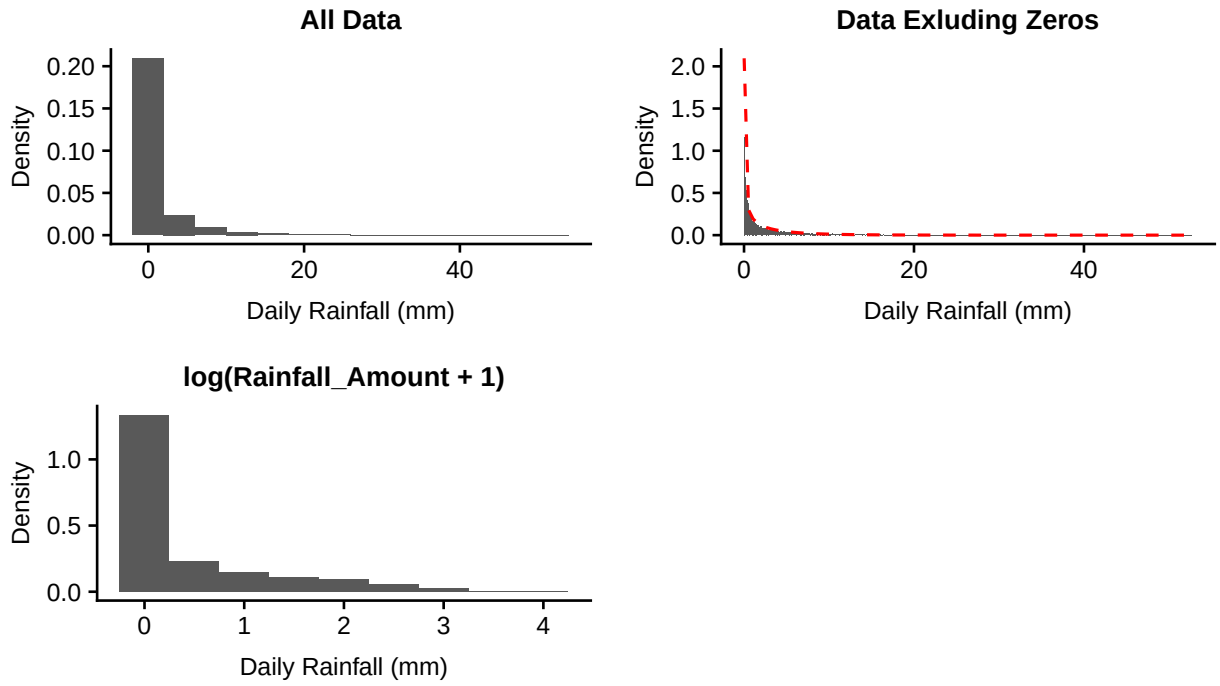


Figure 5: Histograms of daily rainfall amounts using all the data, the data excluding the zeros and the log transform of the data, for the South Mediterranean cluster for all years (1918 to 2017). The red, dashed line in the top right plot indicates a gamma distribution that has been fit to the data using maximum likelihood.

A very high proportion of the data from the South Mediterranean cluster consists of zeros (*Figure 5, top left*) - exactly 18 984 out of the 36 525 days (52.0%) receive no rainfall. It also is seen that the distribution of the daily rainfall is highly skewed to the right. Even after the the data has been logged transformed (*Figure 5, bottom*), the distribution of the response is still highly skew.

The distribution of the response where all the zeros have been removed is seen based on the top right plot of *Figure 5*. A gamma distribution is fit to the data and it appears to fit the data well. This leads to the possibility of using the zero inflated gamma distribution (ZAGA) when modelling rainfall.

Table 5: Five number summary of daily rainfall amount over all years (1918 to 2017) for the South Mediterranean cluster.

Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean
0.0	0.0	0.0	0.76	52.9	1.3

The highly skewed nature of the *cluster daily average rainfall* data is seen on *Table 5*. 52% of the days receive less than 0.76 mm of rainfall. Then, there is a huge jump from the 3rd quartile (0.76 mm) to the maximum (52.9 mm). However, compared to the results of an individual weather station (*Table 3*), it is clear how clustering weather stations and then using the *cluster daily average rainfall* reduces the number of zeros within the data as well as reducing the variability. It is assumed that these properties are similar across all weather stations and clusters.

4.1.3 Visualising Correlations Between Weather Stations

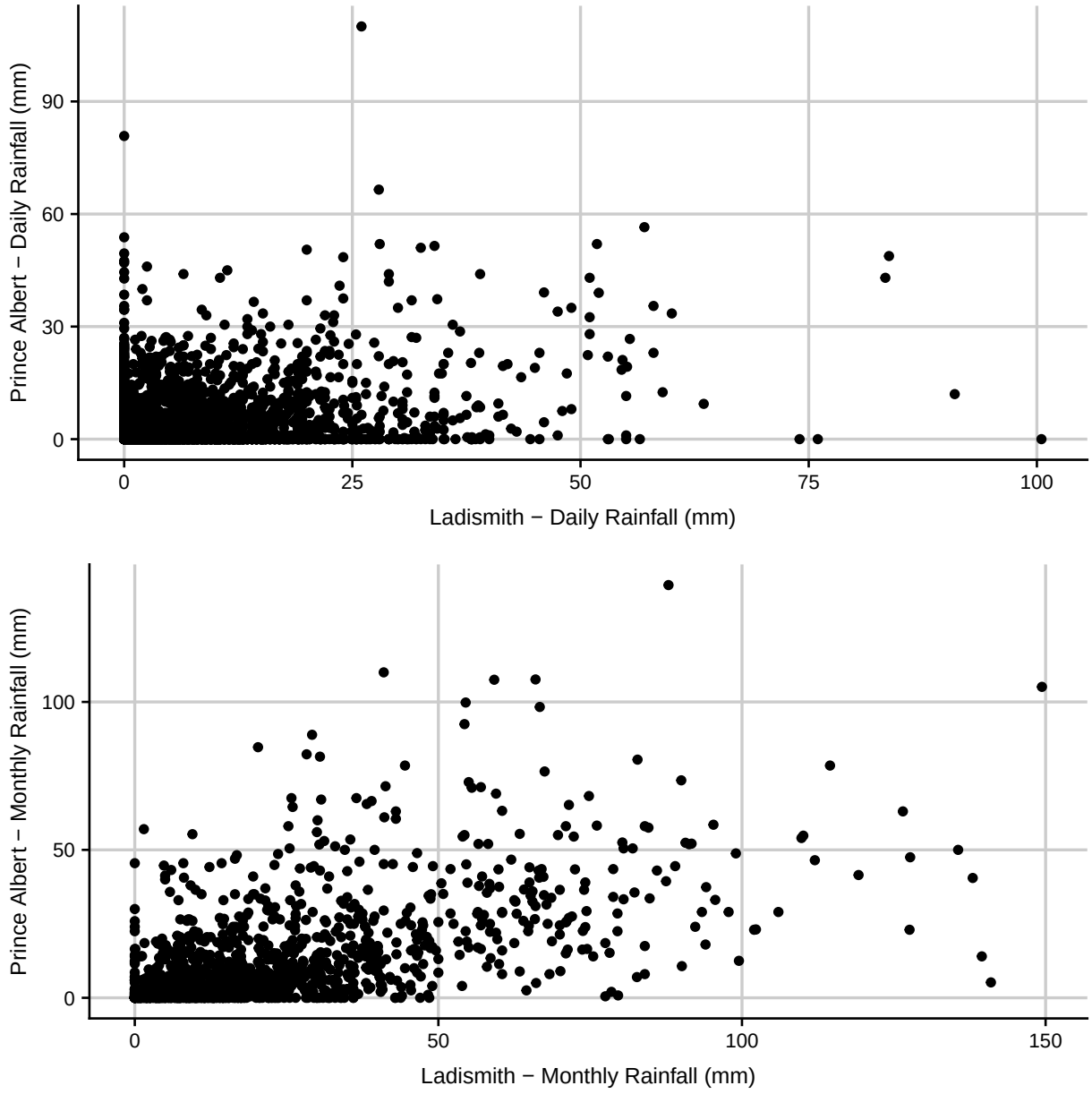


Figure 6: Correlations between weather stations. The top plot shows correlations between the daily rainfall over all years (1918 to 2017) between Prince Albert and Ladismith - the correlation coefficient is 0.42. The bottom plot shows correlations between the monthly rainfall over all years (1918 to 2017) between the same two stations - the correlation coefficient is 0.56.

4.2 Imputing the Missing Years

Annual total rainfall values have been imputed for the years that are missing for certain weather stations. For example, the *Grabouw* weather station is missing years 1956 and 1966 and so values have been imputed for these years. These values are imputed for the purpose of clustering, but are not used in any of the analysis for the start and end of season.

Table 6: Imputed annual total rainfall amounts for 1956 for the Grabouw weather station. Imputed value is indicated in bold.

Year	1954	1955	**1956**	1957	1958	1959
Total Annual Rainfall	1647	1345	**922**	1535	1008	1045

Table 7: Imputed annual total rainfall amounts for 1966 for the Grabouw weather station. Imputed value is indicated in bold.

Year	1964	1965	**1966**	1967	1968
Total Annual Rainfall	1038	1040	**761**	1108	890

These appear to be reasonable estimates for the annual total rainfall (*Tables 6 and 7*). Even though the values are slightly lower than would be expected, this is because the imputation accounts for years of drought and much lower levels of rainfall that have been experienced.

4.3 Clustering

Clustering is performed to group weather stations with similar properties together. Separate analysis is then performed on the final chosen clusters.

4.3.1 Hierarchical Clustering

The 30 weather stations are grouped based on their Euclidean distance using hierarchical clustering. Clustering is performed in order to group together weather stations that have similar rainfall properties. At least three clusters are expected to be formed to account for the three different climatic regions of the Western Cape.

Clustering Using the *Imputed Annual Total Rainfall* Data

First, checking the dendogram to evaluate the optimal number of clusters.

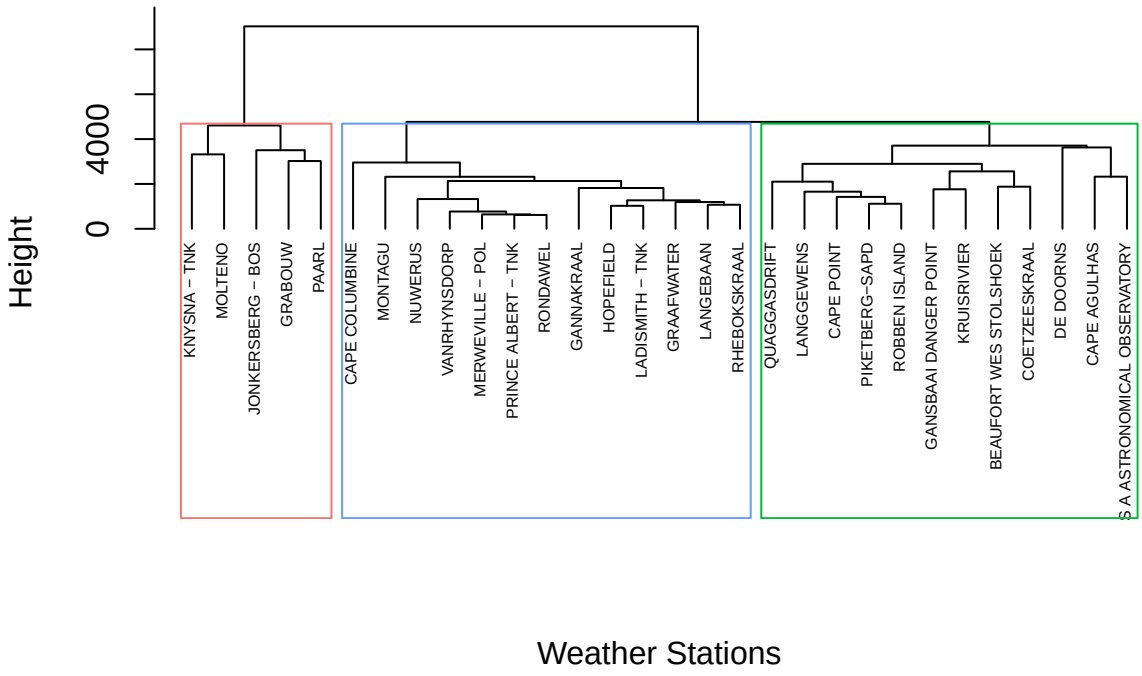


Figure 7: Dendogram plot of hierarchical clustering using the imputed total annual rainfall data.

Based on *Figure 7*, it appears as if the weather stations fall into two distinct groups. This is based on the 'height' which appears on the y-axis of *Figure 7*. However, as there are three known different climate regions within the Western Cape (reference), three clusters are chosen. Looking at these weather stations and where they are located on the map of the Western Cape (*Figure 8*), it shows that they do not fall into the three underlying climatic regions of the Western Cape. But rather, they overlap these regions. This is because weather stations with similar total annual rainfall amounts are being grouped together as opposed to weather stations with similar seasonal trends.

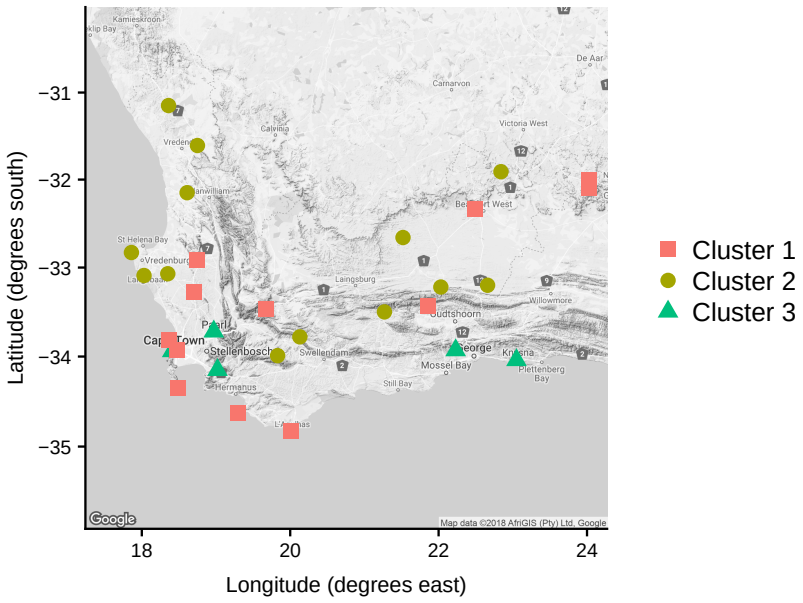


Figure 8: Locations of the weather stations of the three clusters formed using hierarchical clustering based on the imputed annual total rainfall data. Source: Google Maps.

Table 8: Summary of the weather stations belonging to the three clusters formed using hierarchical clustering based on the imputed annual total rainfall data.

Cluster 1	Cluster 2	Cluster 3
Beaufort Wes Stolshoek	Cape Columbine	Grabouw
Cape Agulhas	Gannakraal	Jonkersberg - Bos
Cape Point	Graafwater	Knysna
Coetzeeskraal	Hopefield	Molteno
De Doorns	Ladismith	Paarl
Gansbaai Danger Point	Langebaan	
Kruisrivier	Merweville - Pol	
Langgewens	Montagu	
Piketberg-SAPD	Nuwerus	
Quaggasdrift	Prince Albert	
Robben Island	Rhebokskraal	
SA Astronomical Observatory	Rondawel	
	Vanrhynsdorp	

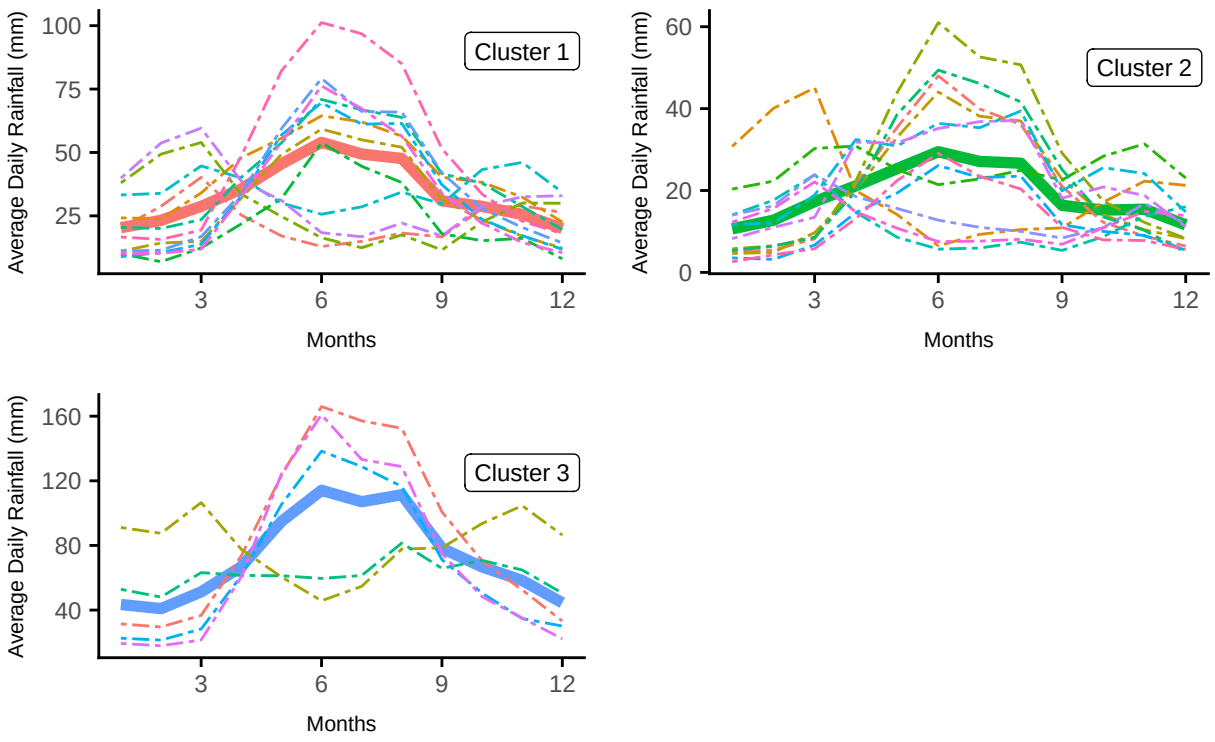


Figure 9: Monthly average rainfall over all years of clusters formed by hierarchical clustering on imputed total annual rainfall data set. The solid line depicts the average monthly rainfall for the whole cluster whilst the dashed lines are the monthly averages of the individual weather stations within the cluster calculated between 1918 to 2017.

The monthly average rainfall amounts of each cluster are seen in *Figure 9*. It is seen that there is no cluster where a predominant rainy summer is captured even though a few of the stations show clear summer rainfall and in this respect are quite different to the predominant winter rainfall. This is a weakness in using the total annual rainfall to form the clusters. Only looking at the total annual rainfall for each year ignores the rainfall patterns throughout the year and thus, the clustering captures no seasonality even though there clearly is.

Clustering Using the *Monthly Average Rainfall* Data

First checking the dendogram to evaluate optimal number of clusters:

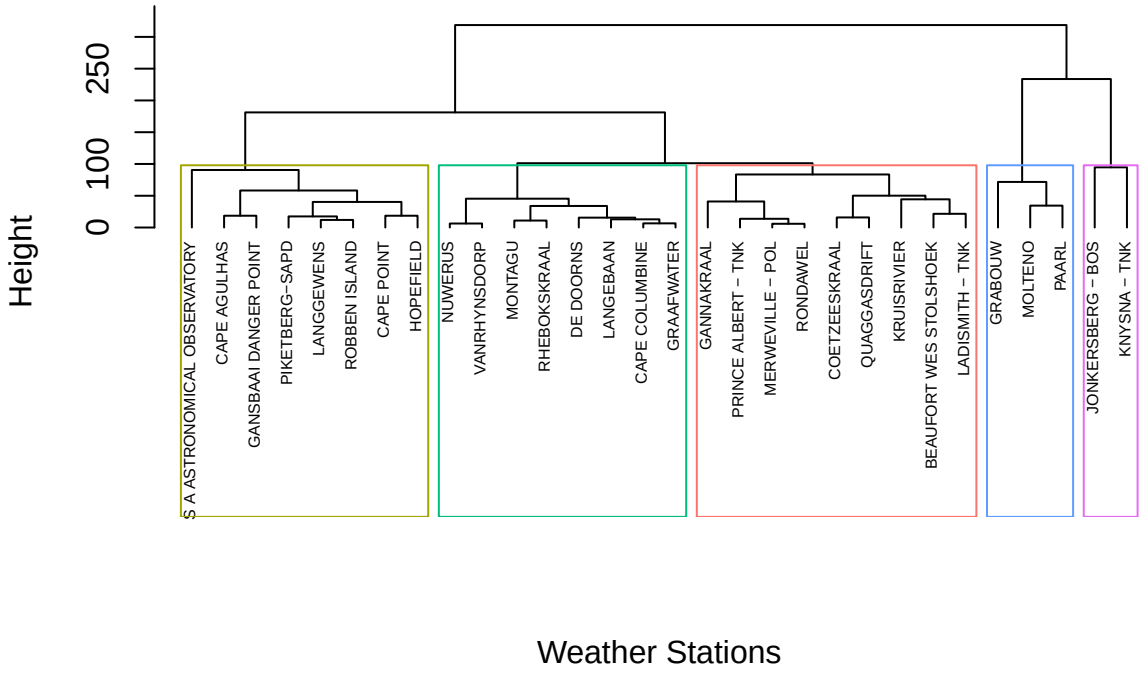


Figure 10: Dendogram plot from hierarchical clustering using average monthly rainfall data.

Even though four clusters should be chosen based on the dendogram plot (*Figure 10*), five clusters are chosen based on visual inspection of the monthly average rainfall plots (*Figure 12*). The addition of the fifth cluster ensures that no summer and winter rainfall weather stations occur within the same cluster.

Comparing these clusters to their location in the Western Cape (*Figure 11*) shows a result that much better captures the seasonal trend of the weather stations. Clusters 2, 3 4 all fall within the Mediterranean region of the Western Cape. The main difference between these three clusters is the total annual rainfall they receive (*Figure 12*). The seasonal trend of these clusters is very similar. The South Coast cluster captures the Karoo region of the Western Cape which receives mostly aseasonal rain with a slight increase during the summer months and the Karoo cluster captures the South Coast region of the Western Cape which also experiences aseasonal rainfall but with a bigger peak during the summer months (*Figure 12*). A big difference between clusters 1 and 5 is on the total annual rainfall they receive as well as the Karoo cluster having a more defined summer rainfall season.

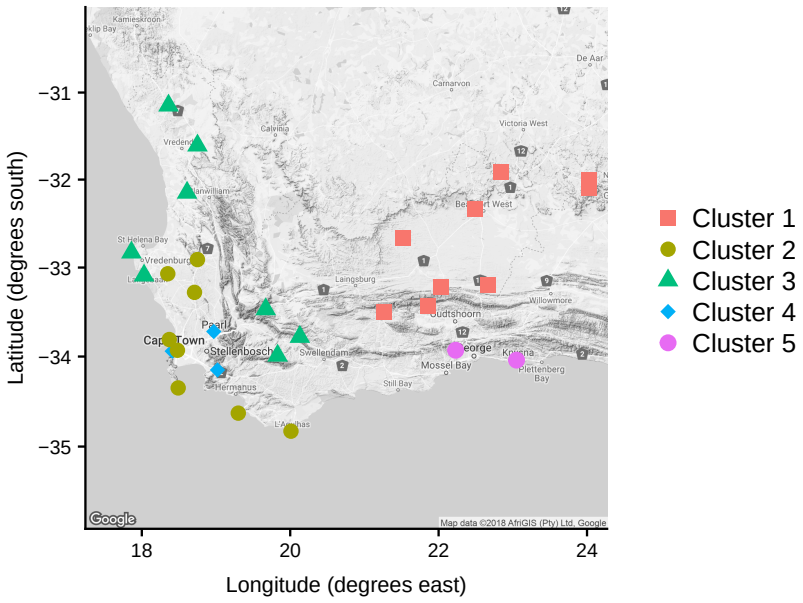


Figure 11: Locations of the weather stations of the five clusters formed using hierarchical clustering based on the average monthly rainfall data. Source: Google Maps.

Table 9: Summary of weather stations belonging to the five clusters formed using hierarchical clustering using the average monthly rainfall data.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Beaufort Wes Stolshoek	Cape Agulhas	Cape Columbine	Grabouw	Jonkersberg - Bos
Gannakraal	Cape Point	De Doorns	Molteno	Knysna
Coetzeeskraal	Gansbaai Danger Point	Graafwater	Paarl	
Kruisrivier	Hopefield	Langebaan		
Ladismith	Langgewens	Montagu		
Merweville - Pol	Piketberg-SAPD	Nuwerus		
Prince Albert	Robben Island	Rhebokskraal		
Quaggasdrift	SA Astronomical Observatory	Vanrhynsdorp		
Rondawel				

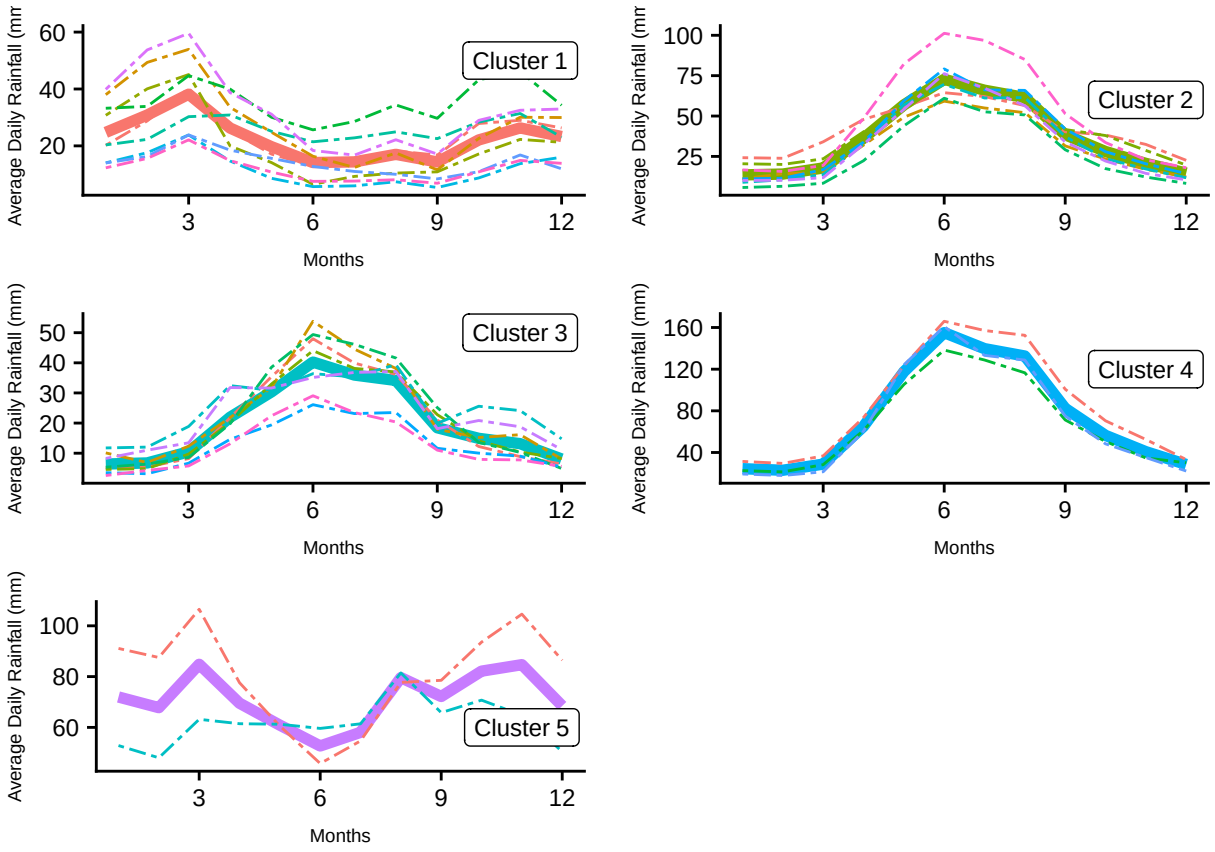


Figure 12: Monthly average rainfall for each hierarchical cluster computed using average monthly data. The solid line depicts the average monthly rainfall for the whole cluster whilst the dashed lines are the monthly averages of the individual weather stations within the cluster calculated between 1918 to 2017. Note the different scales of the y-axes.

It is seen that clustering using the monthly average rainfall per station is much more effective as it captures the actual seasonality of rainfall throughout the year (*Figure 12*). The first cluster and the fifth cluster formed account for the areas where summer rainfall is more prominent. The main difference between these two clusters is not of there seasonal pattern, but their average total annual rainfall. The Karoo cluster receives a higher average total annual rainfall. These clusters are found in the Karoo and South Coast respectively.

Clusters 2, 3 and 4 pick up on the areas that experience winter rainfall, capturing the Mediterranean region of the Western Cape. The main difference between these three clusters is not the seasonal trend throughout the year, but rather the total annual rainfall. As the whole basis of this thesis is to pick up on seasonal trends and changes, these clusters are preferred to the clusters that were formed based only on the total annual rainfall.

4.3.2 K-Means Clustering

The 30 weather stations are grouped based on their Euclidean distance using k-means clustering. Clustering is performed in order to group together weather stations that have similar rainfall properties. At least three clusters are expected to be formed to account for the three different climatic regions of the Western Cape.

Clustering Using the Imputed Annual Total Rainfall Data

Determining the optimal number of clusters based on the *Elbow Method* Plot.

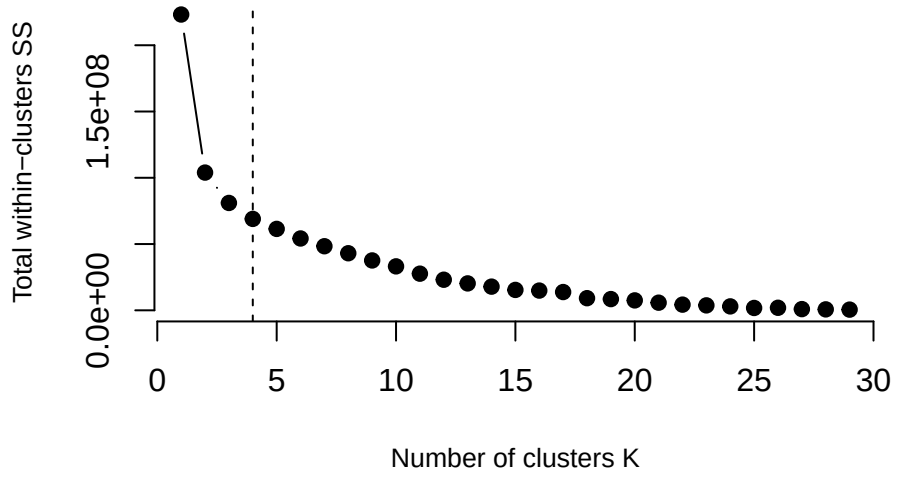


Figure 13: Total within-clusters sum of squares plot. Vertical dashed line indicates chosen number of clusters through visual inspection.

Four clusters are chosen based on *Figure 13*. This is the point at which the total within-clusters sum of squares starts to plateau.

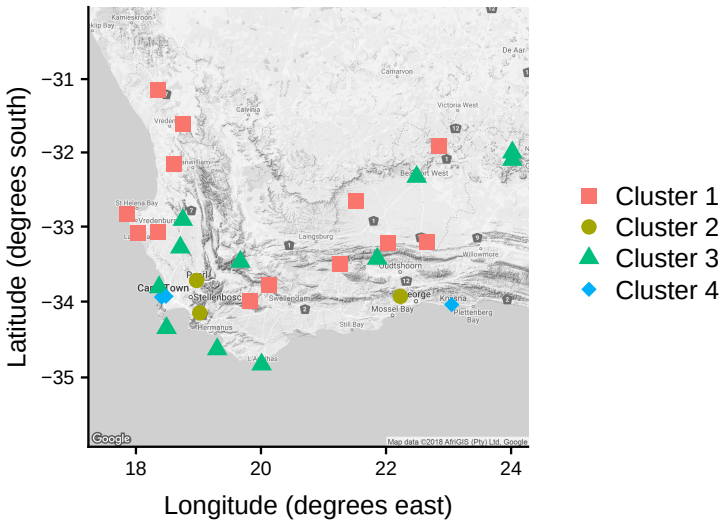


Figure 14: Locations of the weather stations of the four clusters formed using k-means clustering based on the imputed annual total rainfall data. Source: Google Maps.

Table 10: Summary of the weather stations belonging to the four clusters formed using k-means clustering on the imputed annual total rainfall data.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cape Columbine	Grabouw	Beaufort Wes Stolshoek	Knysna
Gannakraal	Jonkersberg - Bos	Cape Agulhas	Molteno
Graafwater	Paarl	Cape Point	SA Astronomical Observatory
Hopefield		Coetzeeskraal	
Ladismith		Gansbaai Danger Point	
Langebaan		Kruisrivier	
Merweville - Pol		Langgewens	
Montagu		Piketberg-SAPD	
Nuwerus		Quaggasdrift	
Prince Albert		Robben Island	
Rhebokskraal		De Doorns	
Rondawel			
Vanrhynsdorp			

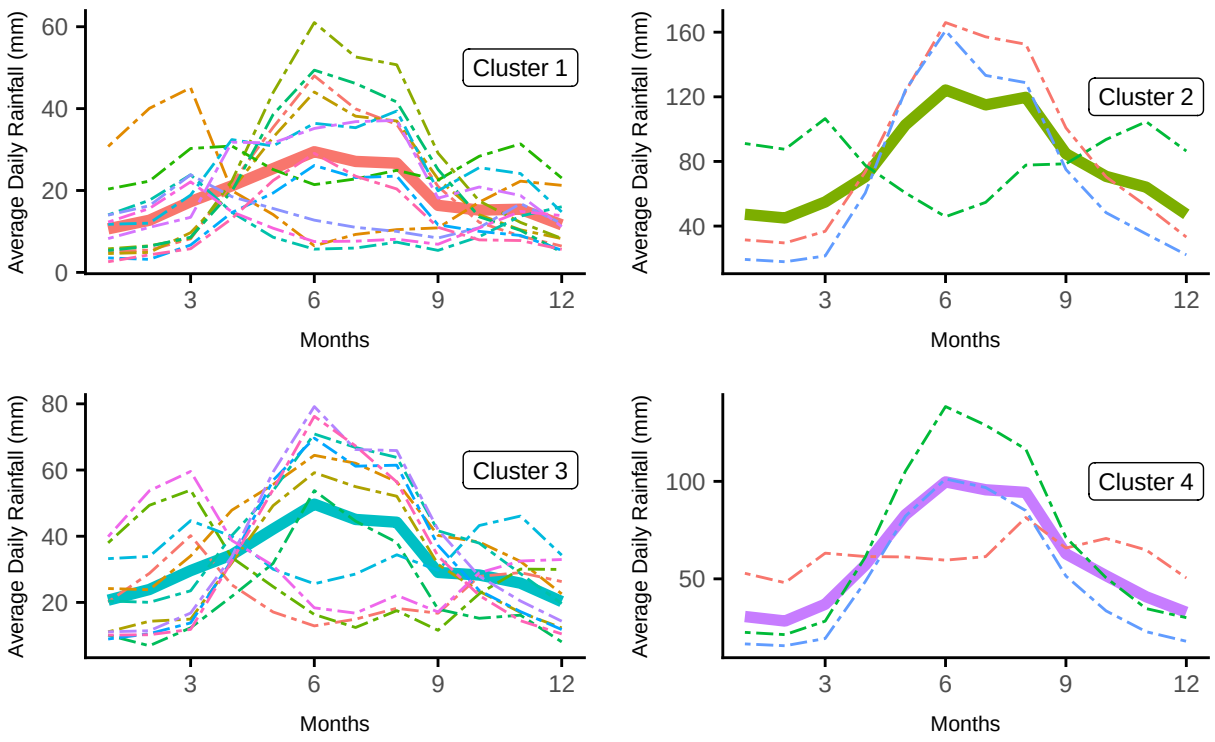


Figure 15: Monthly average rainfall of k-means clusters formed using imputed annual total rainfall data. The solid line indicates the monthly average of the cluster and the dashed lines represent the monthly averages of the individual weather stations calculated between 1918 to 2017. Note the different scales of the y-axes.

The formed clusters capture regions where the annual total rainfall amounts for the year are similar but not the seasonal pattern of the rainfall. As can be seen on *Figure 15*, each cluster has a different amount of rainfall for the year, but some clusters capture both summer and winter rainfall. This method captures no summer rainfall regions when there are known summer rainfall regions within the Western Cape. There are also many weather stations from the three different regions in the Western Cape that are grouped into the same cluster.

This result is very similar to the hierarchical clusters formed on the imputed annual total rainfall. The only difference is that the k-means splits up one of the three hierarchical clusters, forming an extra cluster.

Clustering using the *Monthly Average Rainfall* Data

Determining the optimal number of clusters based on the *Elbow Method* Plot:

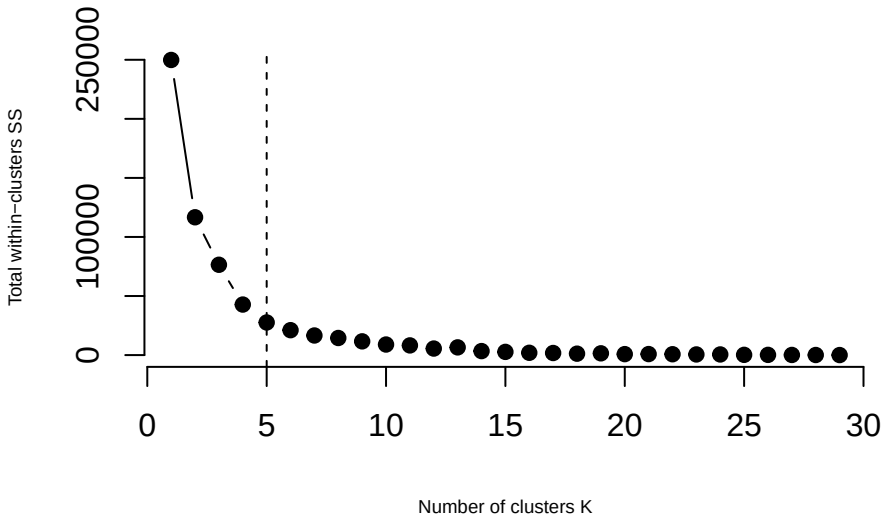


Figure 16: Total within-clusters sum of squares plot. Vertical dashed line indicates chosen number of clusters through visual inspection.

The optimal number of clusters chosen is 5 (*Figure 16*). This is the point at which the total within-clusters sum of squares starts to plateau.

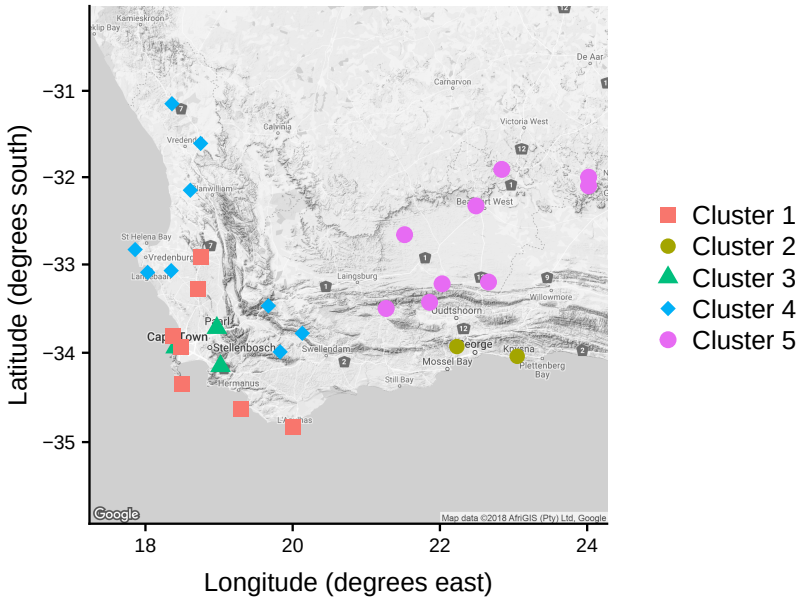


Figure 17: Locations of the weather stations of the five clusters formed using k-means clustering based on the monthly average rainfall data. Source: Google Maps.

Table 11: Summary of the weather stations belonging to the five clusters formed using k-means clustering on the monthly average rainfall data.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cape Agulhas	Jonkersberg - Bos	Grabouw	Cape Columbine	Beaufort Wes Stolshoek
Cape Point	Knysna	Molteno	De Doorns	Coetzeeskraal
Gansbaai Danger Point		Paarl	Graafwater	Gannakraal
Langgewens			Hopefield	Kruisrivier
Piketberg-SAPD			Langebaan	Ladismith
Robben Island			Montagu	Merweville - Pol
SA Astronomical Observatory			Nuwerus	Prince Albert
			Rhebokskraal	Quaggasdrift
			Vanrhynsdorp	Rondawel

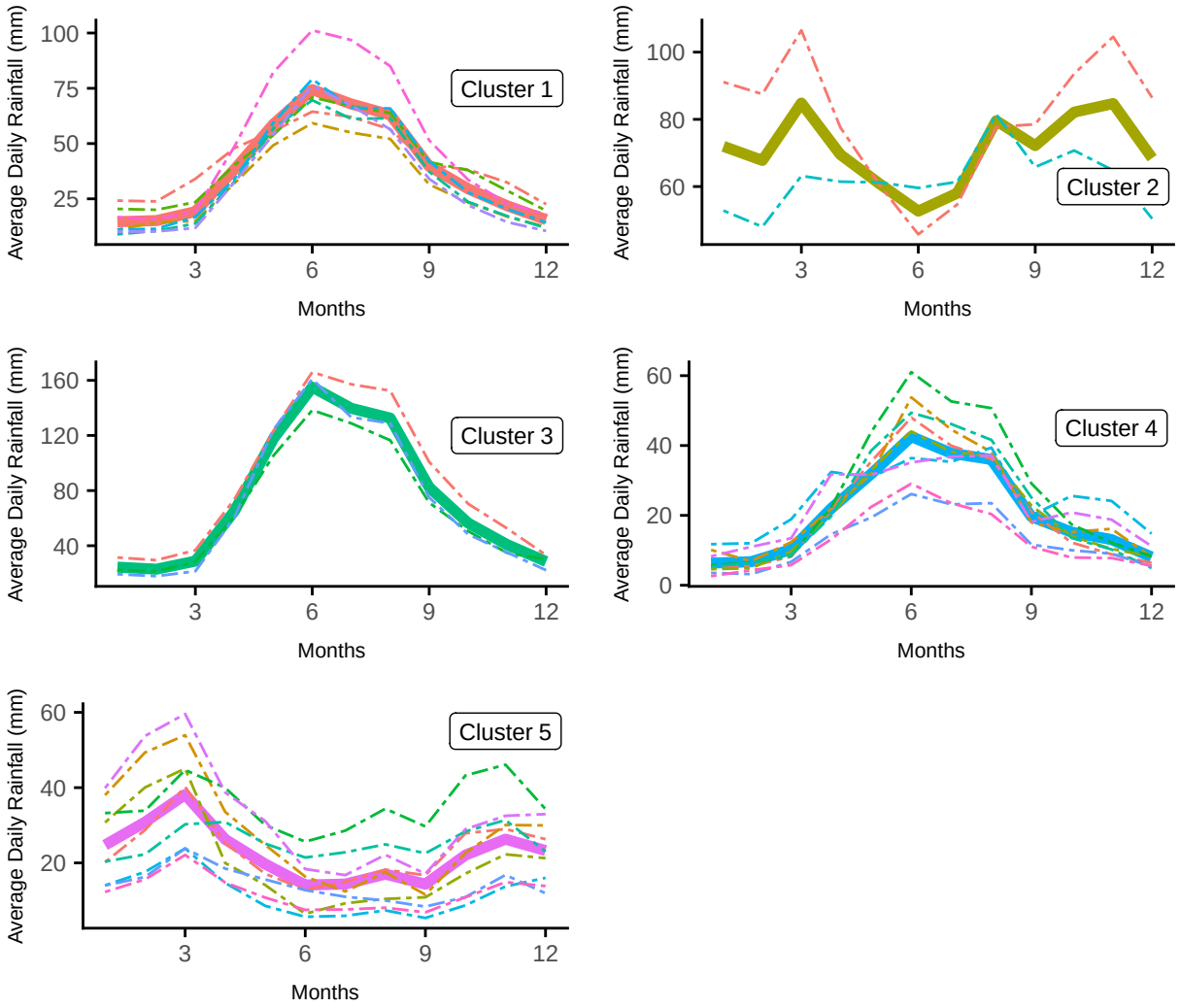


Figure 18: Monthly average rainfall for five K-means cluster formed on average monthly rainfall data. The solid line indicates the monthly average of the cluster and the dashed lines represent the monthly averages of the individual weather stations calculated between 1918 to 2017.

The k-means clusters formed using the monthly average rainfall per weather station capture the seasonal patterns of the weather stations well (*Figure 18*). Within each cluster, the seasonal patterns from the individual weather stations are all very similar. There are clusters where the winter rainfall season is captured (clusters 1,3 and 4) as well as where the summer rainfall season is captured (clusters 2 and 5).

The three regions of the Western Cape are clearly evident in these five clusters that are formed. Clusters 1,3 and 4 are from the Mediterranean region. The only difference between these three clusters is the total annual rainfall that they receive. The seasonal pattern of these three clusters is almost identical. The South Coast cluster is from the South Coast region and the Karoo cluster is from The Karoo. The main differences between clusters 2 and 5 is also the total annual rainfall they experience. The South Coast cluster appears to have a higher total annual rainfall average whilst the Karoo cluster has a more defined summer rainfall season (*Figure 18*).

The advantage of clustering on the monthly average data rather than total annual rainfall is clearly seen with these results. These results are exactly the same as the clusters produced from hierarchical clustering when using the monthly average rainfall data.

4.3.3 Clustering Conclusion

Based on the various clusters formed using different data sets and methods of k-means clustering and hierarchical clustering, the optimal clusters are determined. Five clusters are chosen based on both clustering methods using the average monthly rainfall per station data. These methods both produced the same clusters when using the *monthly average rainfall* per station data. These clusters visually appear to accurately group together weather stations with similar average seasonal patterns for the different regions of the Western Cape. These clusters also fit the underlying climate structure of the Western Cape as none of the clusters overlap different regions. Clusters 2 and 3 do have smaller sample sizes and so interpretation of the results for these clusters should be made with caution.

Table 12: Final five chosen clusters. 7, 2, 3, 9 and 9 weather stations belong to clusters 1, 2, 3, 4, and 5, respectively.

Cluster 1 South Mediterranean	Cluster 2 South Coast	Cluster 3 Central Mediterranean	Cluster 4 West Mediterranean	Cluster 5 Karoo
Cape Agulhas	Jonkersberg - Bos	Grabouw	Cape Columbine	Beaufort Wes Stolshoek
Cape Point	Knysna - TNK	Molteno	De Doorns	Coetzeeskraal
Gansbaai Danger Point		Paarl	Graafwater	Gannakraal
Langgewens			Hopefield	Kruisrivier
Piketberg-SAPD			Langebaan	Ladismith - TNK
Robben Island			Montagu	Merweville - Pol
SA Astronomical Observatory			Nuwerus	Prince Albert - TNK
			Rhebokskraal	Quaggasdrift
			Vanrhynsdorp	Rondawel

Table 12 describes the final clusters that are determined and that are used for the rest of this thesis. These cluster results are exactly the same for both the hierarchical and k-means clustering when using the average monthly rainfall. The *cluster daily average rainfall* is now computed by taking the daily rainfall amount for each day, for every year and then averaging it across all the weather stations within the cluster. This is the data that is used for the rest of the research performed.

4.4 Cumulative and Daily Rainfall Plots

Cumulative and seasonal plots are produced to create a better understanding of the daily rainfall patterns for each cluster.

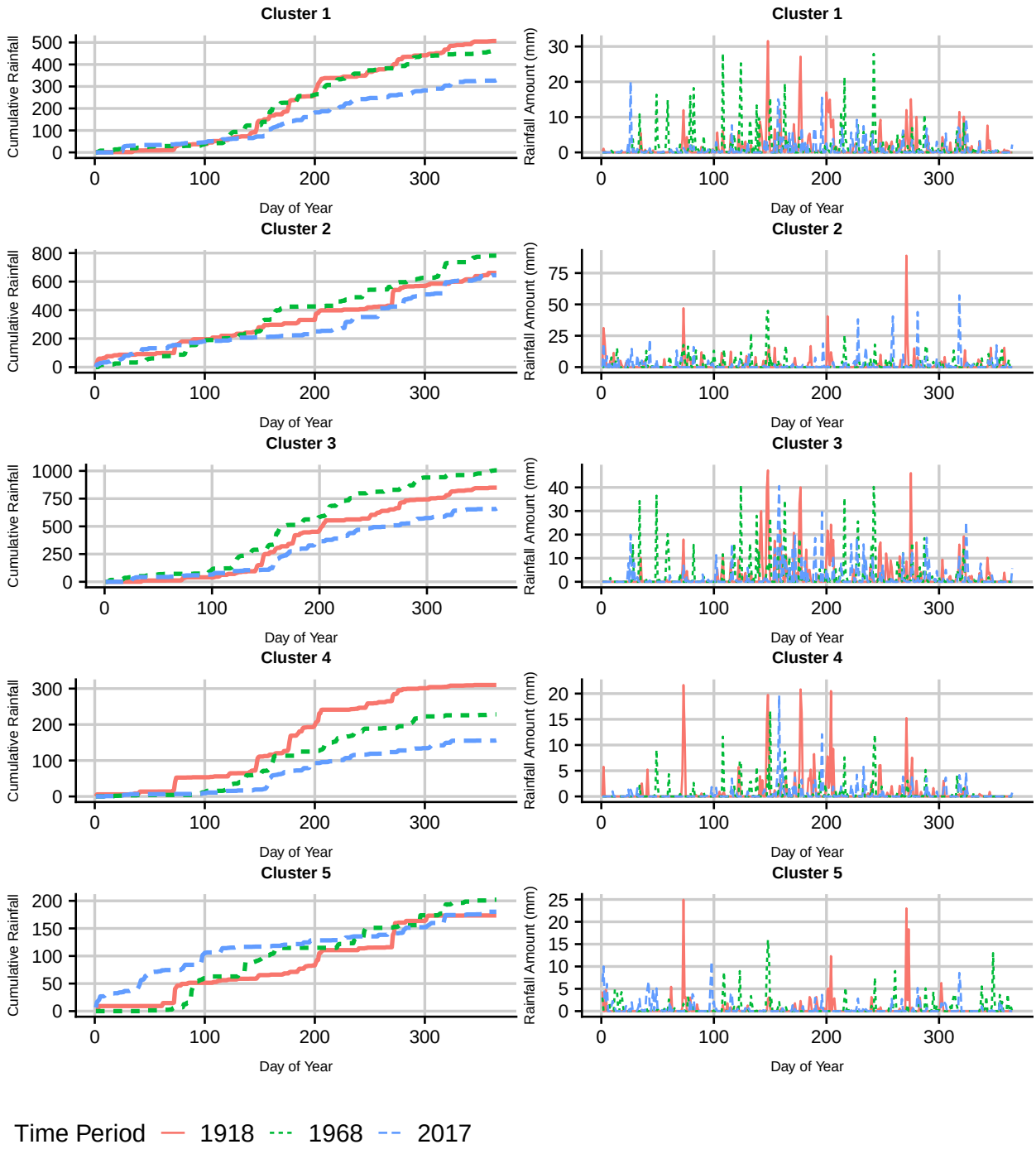


Figure 19: Cumulative plots and annual seasonal profiles of each of the five cluster for three given years (1918, 1968 and 2017). Plots are formed using cluster daily average rainfall data.

The *cluster daily average rainfall* data as defined in *Table 2* is used when computing the above plots (*Figure 19*). Each line represents a year from each cluster.

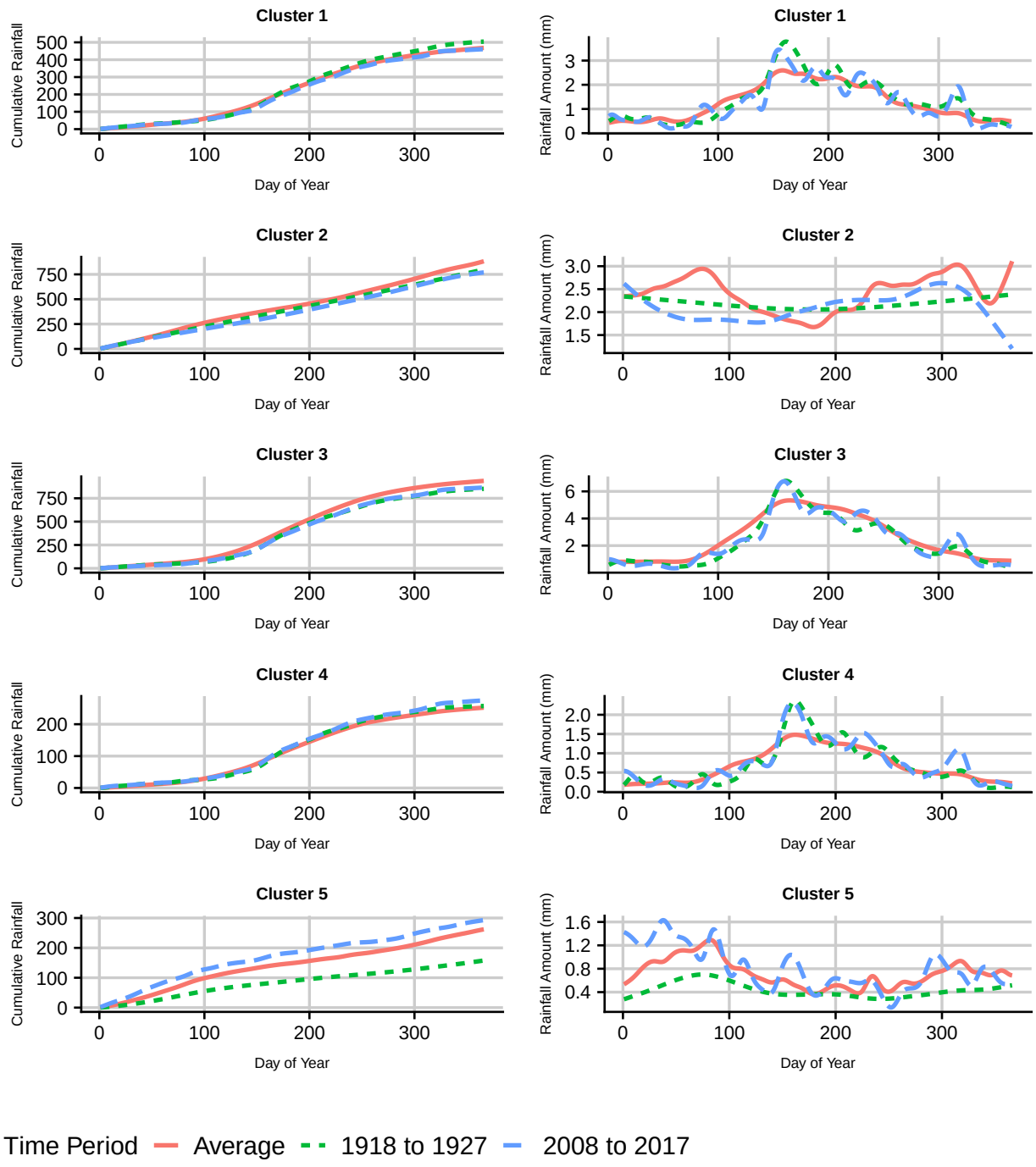


Figure 20: Cumulative and seasonal plots of average for first 10 (1918 to 1927), last 10 (2008 to 2017) and the whole 100 years (1918 to 2017) from each cluster using the cluster daily average rainfall. The average rainfall has been smoothed by fitting splines with a limit of 40 knots.

Summer/Aseasonal Rainfall (South Coast and The Karoo)

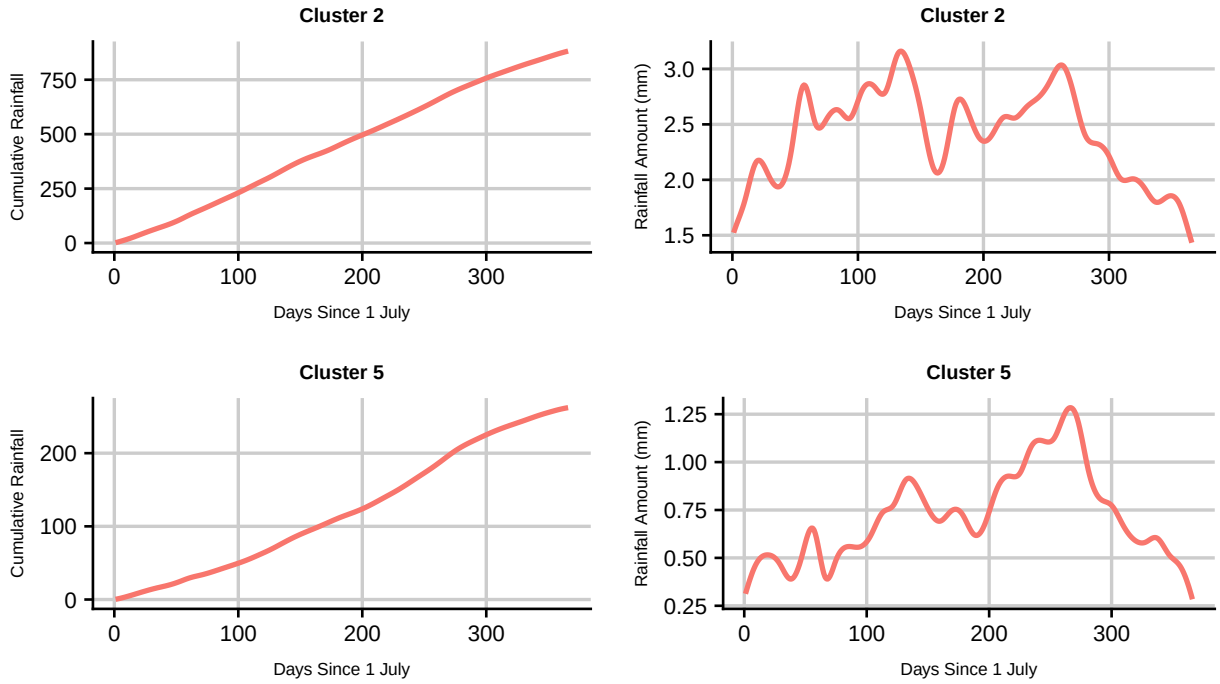


*Figure 21: Cumulative and seasonal plots of average rainfall over all years (1918 to 2017) from summer rainfall clusters. The average of the *cluster daily average rainfall* data has been smoothed by fitting splines. The result is a much smoother profile. The number of knots is limited to 40. The x-axis of these plots starts on 1 July and goes to 31 June of the following year.*

The South Coast cluster appears to have two peaks during the summer months whereas the Karoo cluster appears to have quite a defined peak during the second half of the summer months (*Figure 21*).

4.4.1 Estimating the Annual Cumulative Thresholds for Start and End of Season

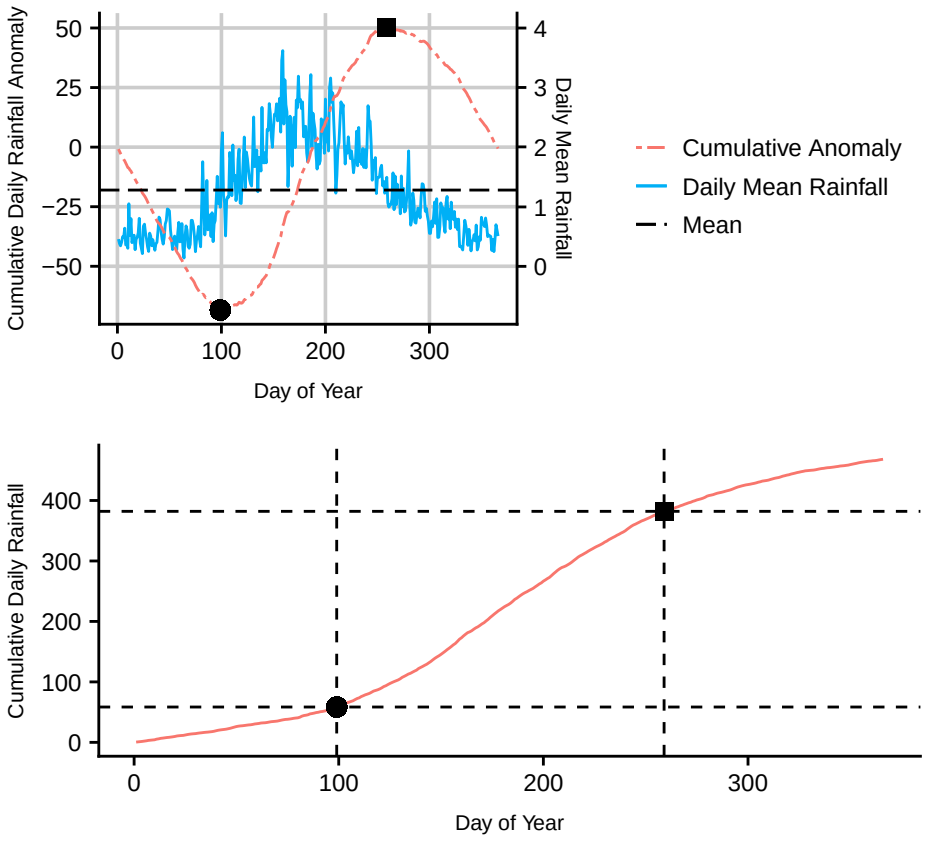


Figure 22: An example of finding the chosen start and end of season thresholds for South Mediterranean cluster using the cluster daily average rainfall. Thresholds are found using the cumulative daily rainfall anomaly. The black points indicate the start and end of season. Thresholds for start and end of season for the South Mediterranean cluster are 58.4 mm and 382.0 mm, respectively.

The thresholds for start and end of season are found using the above method for all clusters. (*Figure 22, bottom*)

Table 13: Annual cumulative thresholds for estimating start and end of season. Thresholds are found using the cumulative daily rainfall anomaly. All values are in millimeters (mm).

Cluster Name	Start of Season	End of Season
South Mediterranean	58.4	382. 0
South Coast	234.1	515.2
Central Mediterranean	111.5	791.1
West Mediterranean	25.3	204.3
The Karoo	115.5	204.2

4.5 Start of Season Plots For the WINTER RAINFALL Clusters

Different methods are used to estimate the start of season so as to determine whether there have been any changes in seasonality. The data that is used in the following methods is based on the *cluster daily average rainfall* data (*Table 2*). The following plots are only for the clusters that experience southern hemisphere winter rainfall (all the Mediterranean clusters (1, 3 and 4)) for all years (1918 to 2017).

4.5.1 Start of Season Estimated Using a Threshold for Annual Cumulative Rainfall

Different thresholds are found based on the cumulative daily rainfall anomaly (*Figure 22*). The thresholds for the South Mediterranean (1), Central Mediterranean (3) and West Mediterranean (4) clusters are 58.4 mm, 111.5 mm and 25.3 mm, respectively (*Table 13*). There is no restriction put on the time period for when the start of season can occur within the year.

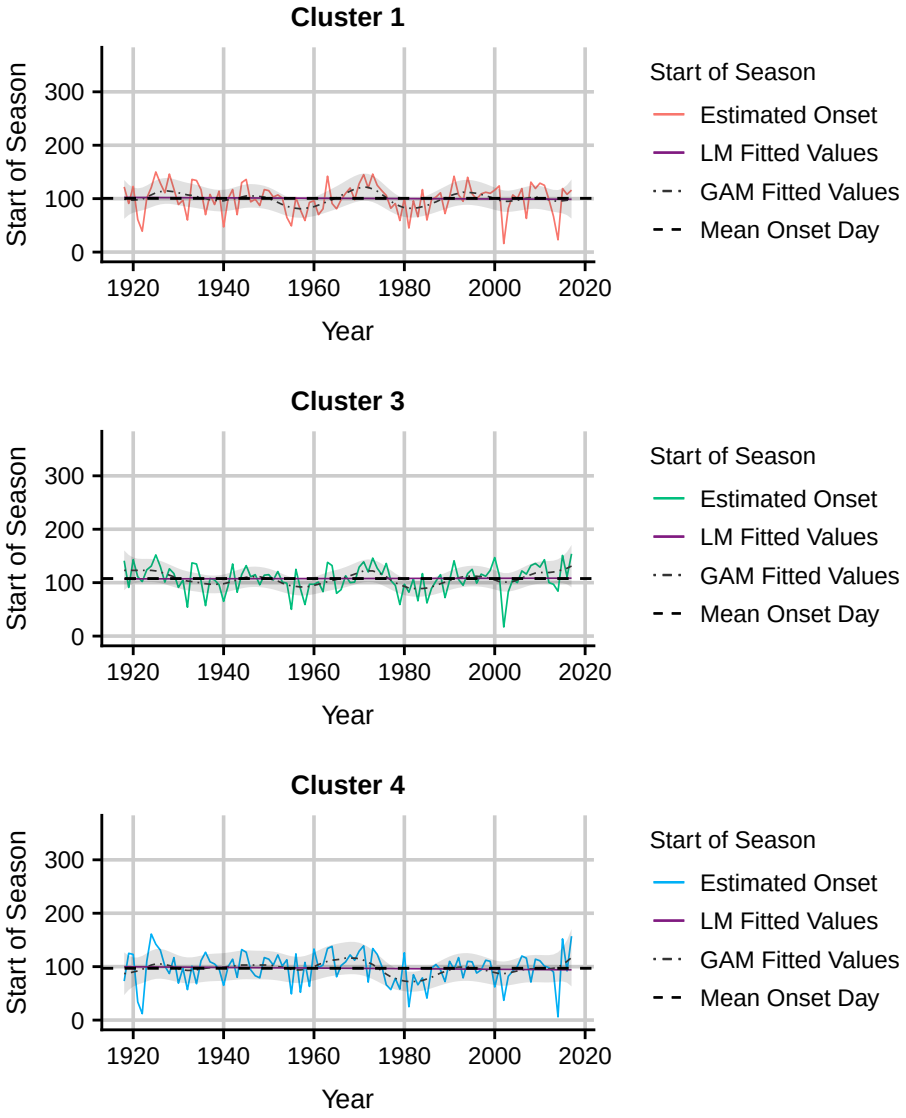


Figure 23: Start of season using a cumulative threshold (1918 to 2017). Thresholds for the South Mediterranean, Central Mediterranean and West Mediterranean clusters are 58.4 mm, 111.5 mm and 25.3 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 14: Summary of the linear model fitted to estimated onset dates using a cumulative threshold. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	101	27.5	-0.03	0.10	0.72	100
Central Mediterranean	108	25.1	0.01	0.09	0.87	100
West Mediterranean	97	29.9	-0.06	0.10	0.58	100

There is no linear trend evident for any of the clusters (*Table 14*) however, all the fitted GAMs show periods of increasing and decreasing trends. For the years 1950 to 2000, the peaks and troughs from all the three clusters appear to follow a very similar pattern (*Figure 23*).

Looking at the non-linear trends:

For the South Mediterranean cluster (1), between 1955 and 1959, the 95% confidence interval of the GAM is completely below the mean onset day. The mean onset day is on the bottom of the boundary for 1971. Between 1980 and 1983, the mean onset day is completely above the confidence interval. Thereafter, the onset day remains stable (*Figure 23, top*).

For the Central Mediterranean cluster (3), the confidence interval is completely below the mean onset day between 1957 and 1958. Between 1981 and 1985, the confidence interval is completely below the mean onset day. For the rest of the time period, the mean onset day remains near the middle of the confidence showing a stable onset day (*Figure 23, middle*).

The mean onset day always falls within the GAM confidence interval for the West Mediterranean cluster (4) except between 1979 and 1984 where the confidence interval shifts completely below the mean onset day (*Figure 23, bottom*).

4.5.2 Start of Season Using Stern et. al. (1981) Definition

The start of season estimated using the Stern et. al. Criteria as defined in (Chapter 3.5.2)

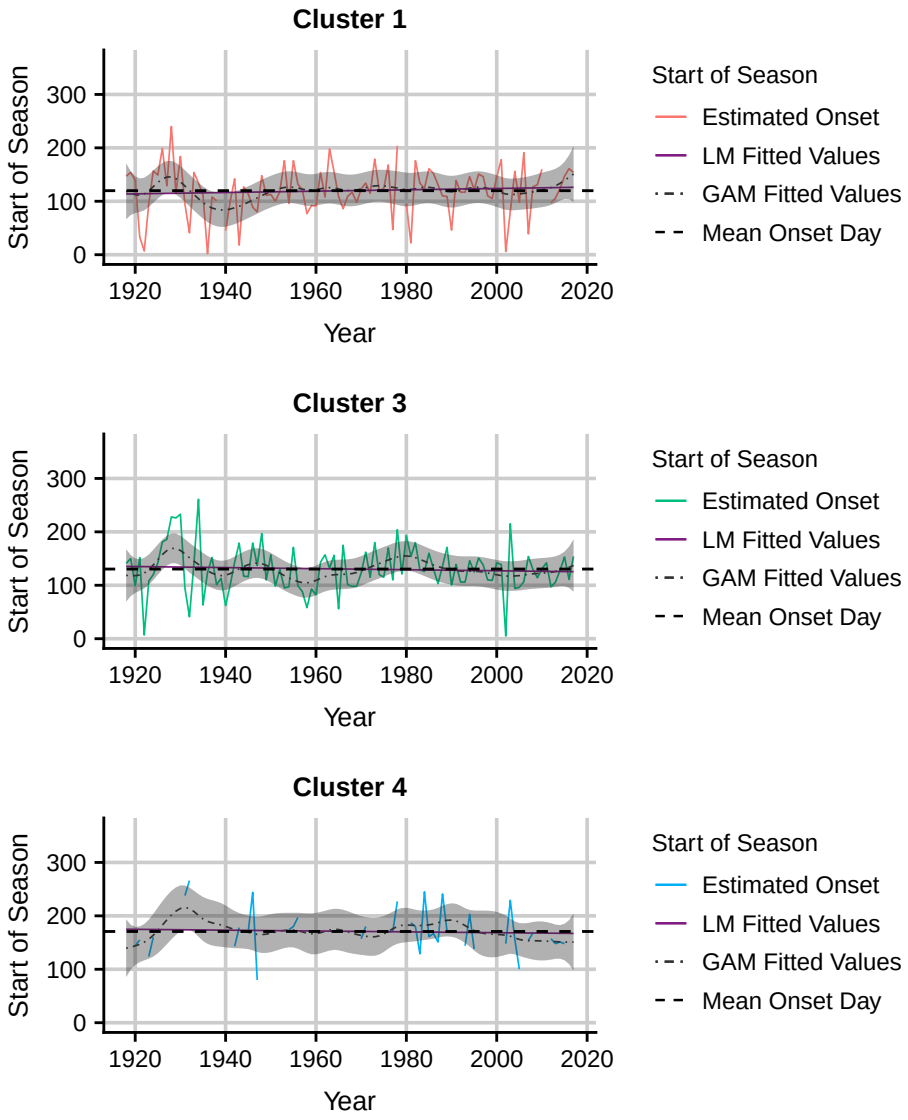


Figure 24: Start of season using Stern et. al. (1981) criteria. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Using Stern et. al.'s definition there are many years where the start of season could not be found as not all the criteria are met. The main reason for this is the constraint that 25 mm of rainfall needs to be captured within a 5 day period. The West Mediterranean does not have high enough levels of rainfall throughout the year and this constraint is often not met. Even though this method is not further analysed, the peak around 1930 that all the clusters experience is noted. These criteria could be used for the South Mediterranean and central Mediterranean as many of the years do meet the requirements however, this method is not further analysed as a method where the criteria have been adjusted is preferred (*Figure 24*).

4.5.3 Start of Season Using Modified Criteria

Certain criteria are to be met for each cluster in order for the start of season to be estimated. Criteria are found in chapter 3.5.2.

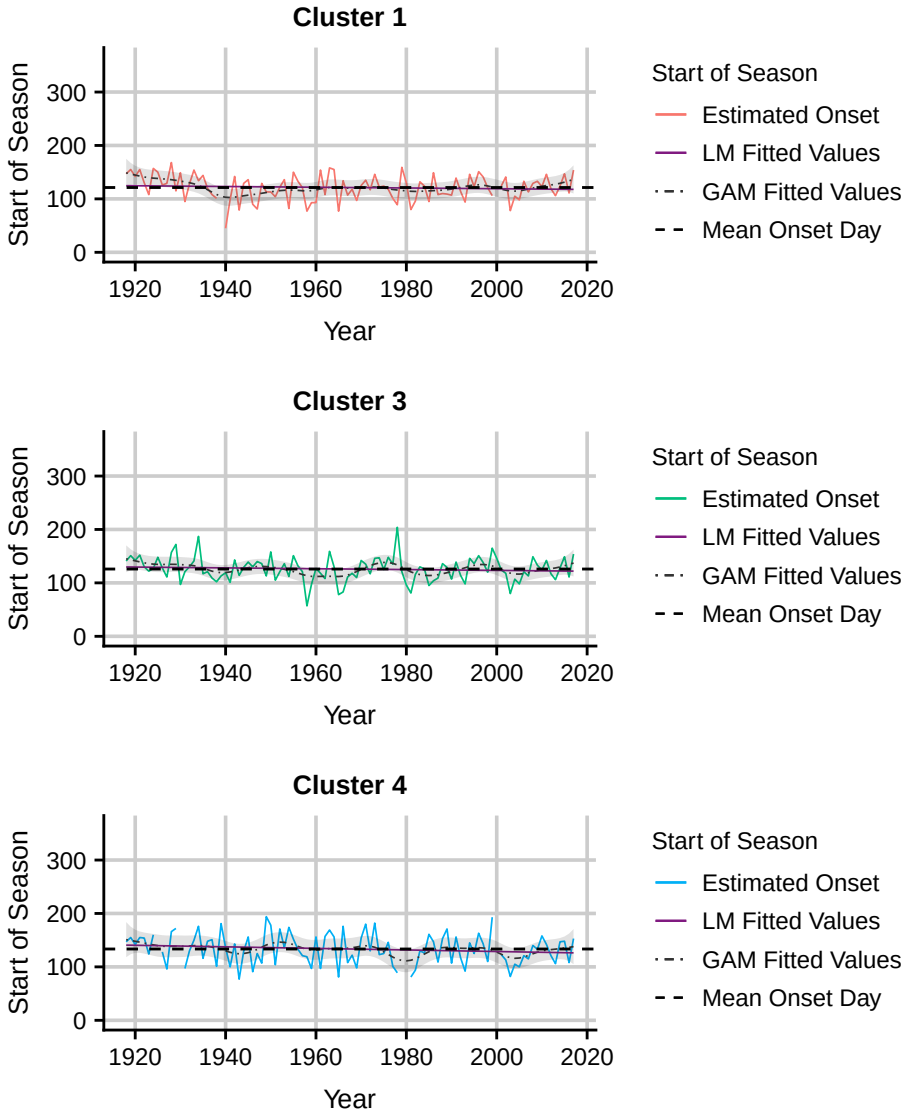


Figure 25: Start of season estimated using a modified criteria (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 15: Summary of the linear model fitted the to estimated onset dates using modified criteria. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	121	23.7	-0.06	0.08	0.41	98
Central Mediterranean	126	22.8	-0.08	0.08	0.33	100
West Mediterranean	133	28.6	-0.15	0.10	0.15	95

Based on the linear models for all of the clusters, none of the gradients show any evidence of linear trend (*Table 15*).

Looking at the non-linear trend of the clusters:

For the South Mediterranean cluster, there seems to be a change to an earlier onset of the rainfall season until 1940. This is supported by the GAM 95% confidence interval being completely above the mean onset day up until 1926. At 1940, the confidence interval then shifts down until it is completely below the mean onset day until 1945 . For the rest of the period, the onset day is really stable (*Figure 25, top*).

For the Central Mediterranean cluster, the mean onset day always falls within the GAM 95% confidence interval. From 1959 to 1966 to the mean onset day is right on the top boundary of the confidence interval and between 1984/85 it also right on the top. Around 1975, the confidence interval shifts upward until the mean onset day is right on the bottom boundary (*Figure 25, middle*).

For the West Mediterranean cluster, the mean onset day also never falls outside the boundary of the GAM 95% confidence interval. However, it does fall on the top of the boundary between 1980 to 1981 and 2004 (*Figure 25, bottom*).

An Example of a GAM - Modelling the Mean Daily Rainfall

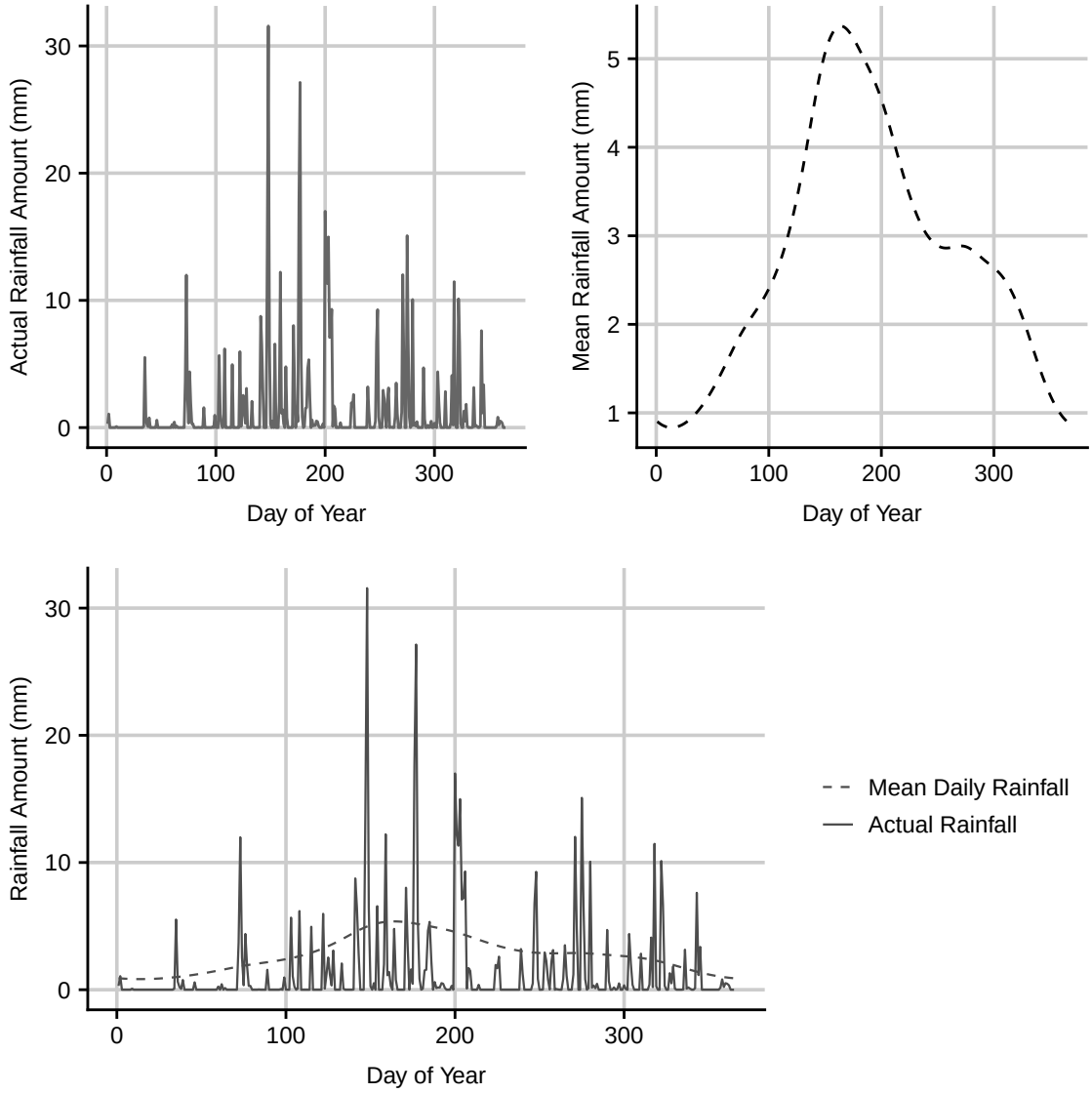


Figure 26: An example of a GAM fitted to the South Mediterranean cluster, 1918. The data that is used is the average of the daily rainfall for the weather stations within that cluster - the cluster daily average rainfall. This GAM is fit using the Zero Adjusted Gamma Distribution due to high proportion of zeros in the data. The GAM is limited to 6 knots and is fit using a cyclical spline basis.

Figure 26 shows how the modeled mean daily rainfall produced from a GAM fits to the actual data. The smooth nature of the GAM is evident. Six knots are chosen - this number is chosen as an increase of knots leads to over fitting of the model and the model becomes too sensitive to outlying events. So, the limit on the number of knots for the GAM deals well with outlying events. Based on the plot of the GAM mean daily rainfall, the potential of using a threshold is clear i.e. the start of the rainfall season could be declared when the GAM moves above 3 mm of rainfall. This should produce better results compared to just looking at a threshold of the raw cumulative data as the GAM should get rid of any outlying events.

4.5.4 Start of Season Using a Threshold Based on GAM Mean Daily Rainfall

Thresholds for the South Mediterranean and Central Mediterranean clusters are 2.75 mm and 7mm, respectively. For both of these clusters the search for the start of season is limited to days between the 50th and 185th day of the year so as to exclude any outlying events.

For the West Mediterranean cluster, a threshold of 2.2 mm is chosen and the search is limited to days after the 40th and 185th day of the year.

The thresholds are based on approximately the maximum rainfall of the average smooth profile (*Figure 20*). The reason the maximum is chosen is because the plots show the average profile as apposed to the actual profile. The thresholds were also chosen to ensure that a high number of the years exceed the threshold.

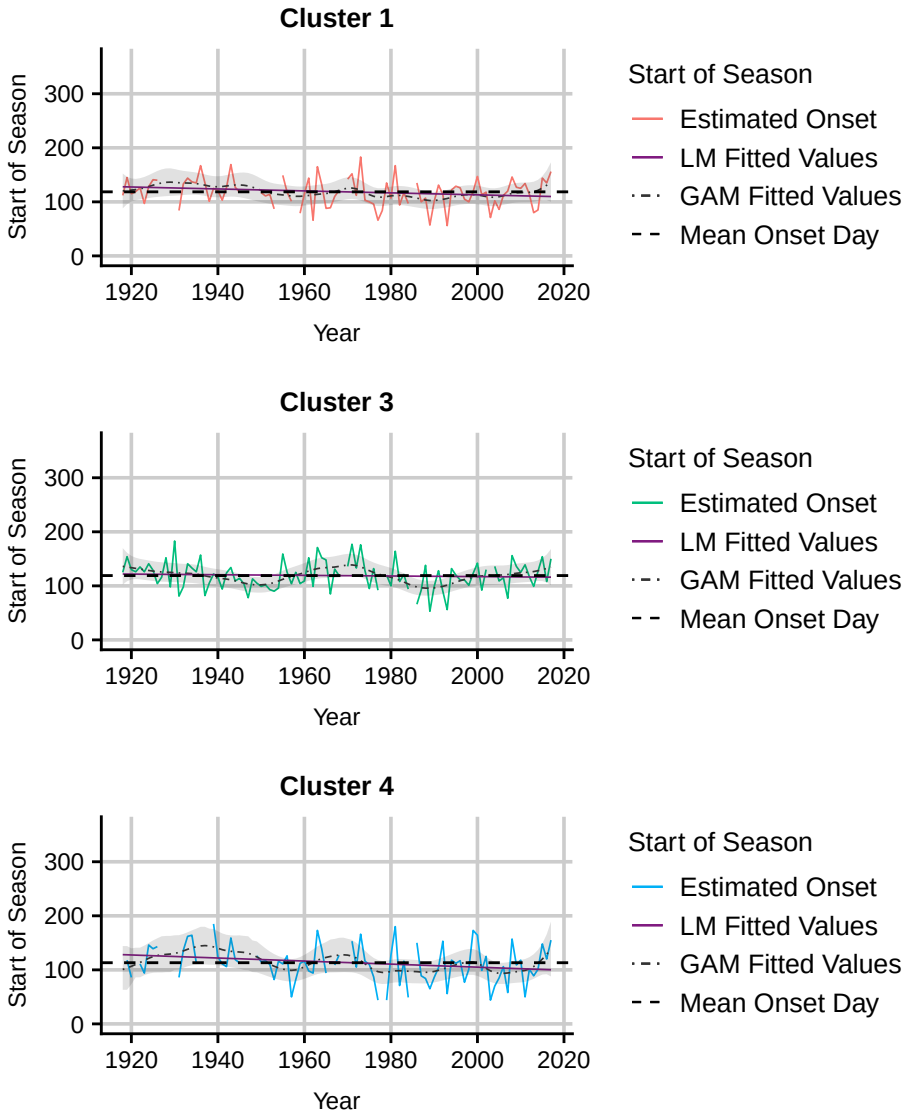


Figure 27: Start of season estimated using a threshold from GAM mean daily rainfall (1918 to 2017). The thresholds for the South Mediterranean, Central Mediterranean and West Mediterranean clusters are 2.75 mm, 7 mm and 2.2 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 16: Summary of the linear model fitted to the estimated onset dates using a threshold based on the GAM mean daily rainfall. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	119	26.6	-0.18	0.09	0.06	90
Central Mediterranean	119	25.9	-0.06	0.09	0.53	96
West Mediterranean	113	34.6	-0.28	0.13	0.03	86

For the South Mediterranean cluster, there is marginal evidence that the start of season day is shifting earlier in the year based on the linear model and its p-value of 0.06 (*Table 16*). According to the model, the onset of the season is shifting 0.18 days earlier each year. Looking at the non-linear trend, the GAM 95% confidence interval shifts downward from 1918 until 1989 where the confidence interval is completely below the mean onset day between 1989 and 1990. This supports the linear model's results suggesting an earlier start in the rainfall season up until approximately 1990 (*Figure 27, top*).

The Central Mediterranean cluster shows no evidence of linear trend over time (*Table 16*). There is evidence of non-linear trend. The GAM 95% confidence interval shifts completely below the mean onset day between 1947 and 1952 as well as between 1986 and 1994. Around 1971 the confidence interval shifts upward until the mean onset day is right on the bottom boundary. In the last 20 years, the confidence interval is shifting upward so that the mean onset day is getting closer to the bottom of the confidence interval pointing toward a later start to the rainfall season (*Figure 27, middle*).

The West Mediterranean cluster shows marginal evidence of a decreasing linear trend with a p-value of 0.03 (*Table 16*). The onset day seems to be shifting 0.28 days earlier per year. Even though the GAM 95% confidence interval seems to be shifting downward, the mean onset day is always found within it. Around 1937, the confidence interval has shifted upward and the confidence interval is near the bottom of the boundary. The confidence interval then starts shifting downward until the mean onset day passes near the top boundary around 1988. The mean onset day also passes near the top boundary around 2006 (*Figure 27, bottom*).

4.5.5 Start of Season Based on the Gradient of the GAM Mean Daily Rainfall

The gradient of the mean daily rainfall is examined to estimate the start of season. Similar to previous methods, the search is limited between the 50th and 185th day of the year for all the clusters so as to exclude any outlying events.

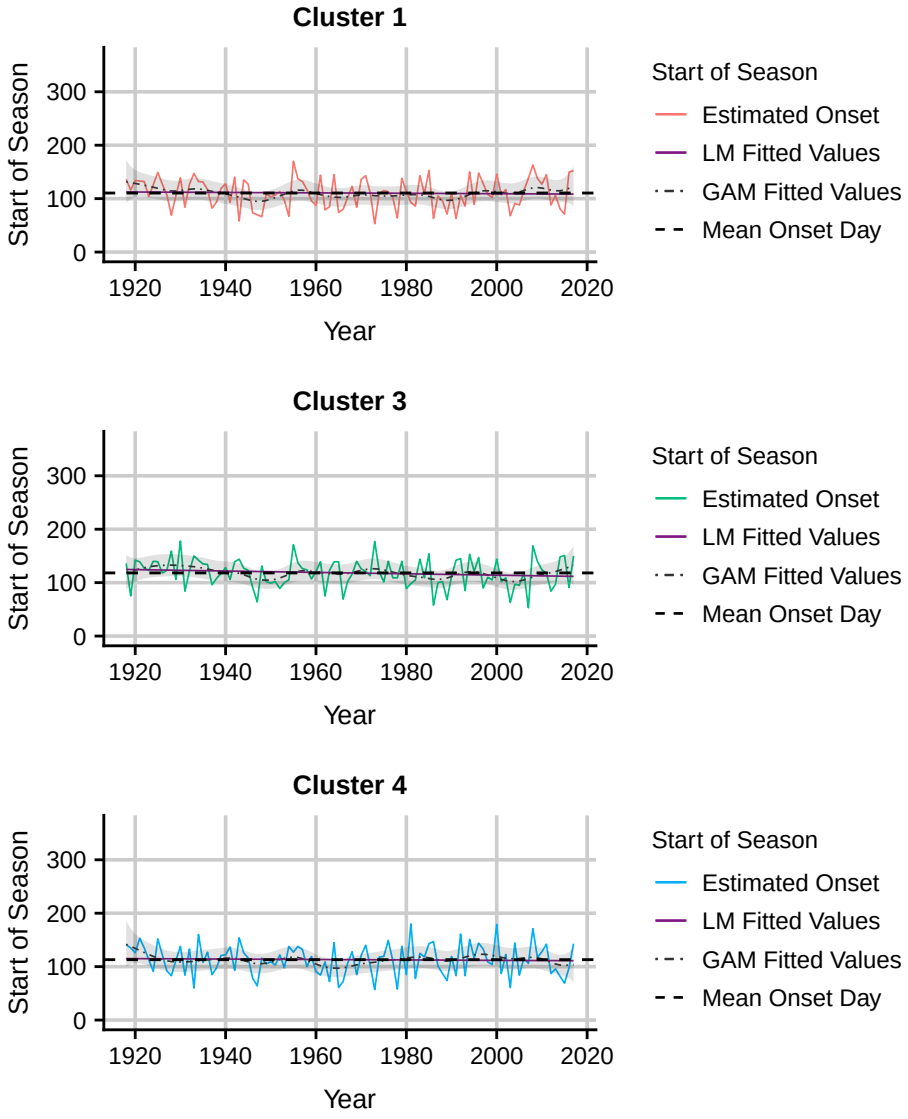


Figure 28: Start of season estimated using the gradient of the GAM mean daily rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 17: Summary of the linear model fitted to the estimated onset dates using the gradient of the GAM mean daily rainfall. 'n' refers to the of number years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	111	27.5	-0.04	0.10	0.68	100
Central Mediterranean	118	26.4	-0.13	0.09	0.16	100
West Mediterranean	113	28.4	-0.04	0.10	0.69	100

All the Mediterranean clusters show no evidence of linear trend (*Table 17*).

Looking at the non-linear trend of the clusters:

For the South Mediterranean cluster, initially, the GAM 95% confidence interval is quite elevated. It then starts to shift downward until it is completely below the mean onset day at 1947. Around 1989, the mean onset day passes through the top boundary as the confidence interval has shifted downward (*Figure 28, top*).

Initially for the Central Mediterranean cluster, the GAM 95% confidence interval is raised. It then shifts downward such that the mean onset day passes near the top boundary around 1949. Around 1987, the mean onset day passes through the top boundary. Lastly, around 2004, the confidence interval shifts completely below the mean onset day (*Figure 28, middle*).

For the West Mediterranean cluster, even though the GAM has quite a few bumps, they are all small and mean onset day does not leave the confidence interval. The confidence interval does shift down and the mean onset day passes near the top of the interval around 1965. Thereafter, the onset day is fairly stable (*Figure 28, bottom*).

An Example of a GAM - Modelling the Probability of Zero Rainfall

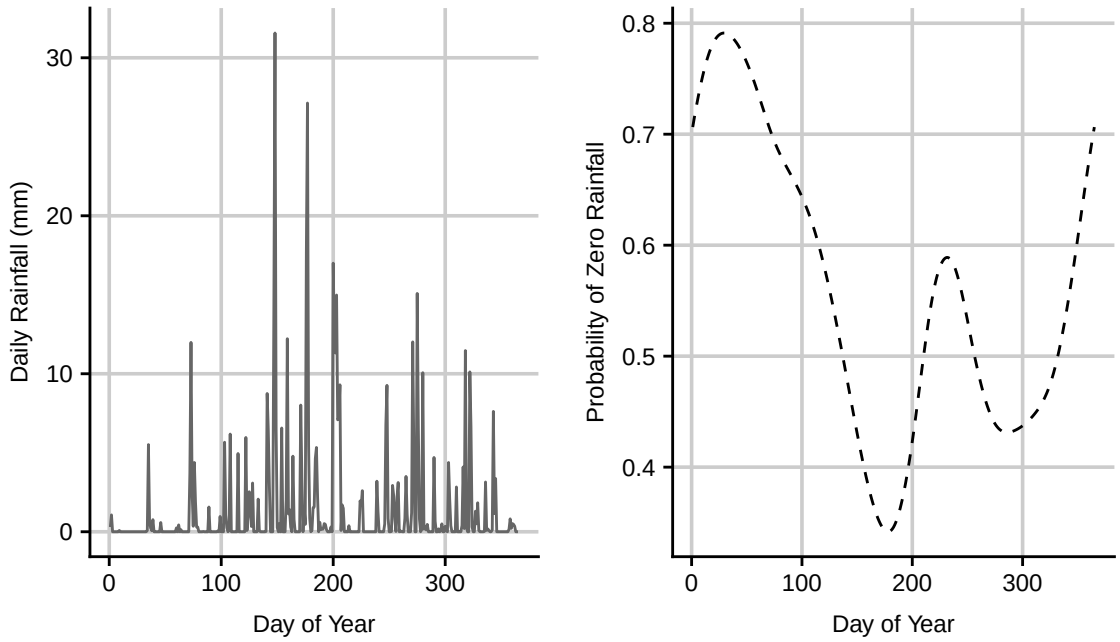


Figure 29: An example of a GAM for a given year (1918) modelling the probability of zero rainfall for the South Mediterranean cluster. The modelling is done using the cluster daily average rainfall. Modelling is performed using the zero adjusted gamma distribution.

As part of the Zero Adjusted Gamma distribution, the probability of zero rainfall for each day can be modeled. The probability of zero rainfall decreases toward the winter months where winter rainfall is experienced i.e. there is a lower probability of zero rainfall being captured in the winter months - it is more likely to rain (*Figure 29, right*). As this results in a fairly smooth curve, the potential advantage of using a threshold of the probability of zero rainfall is evident. For example, a threshold of 0.5 can be chosen. Once this value has been crossed, the onset day can be estimated.

4.5.6 Start of Season Using a Threshold Based on the Probability of Zero Rainfall

Different thresholds are used for different clusters as well as limiting the start of season to a certain time period within the year. The chosen thresholds for the South Mediterranean, Central Mediterranean and West Mediterranean clusters are 0.5, 0.65 and 0.65, respectively. The search is limited between the 50th and 185th day of the year so as to exclude any outlying events. These thresholds are chosen based on a visual inspection of the modeled probability of zero rainfall.

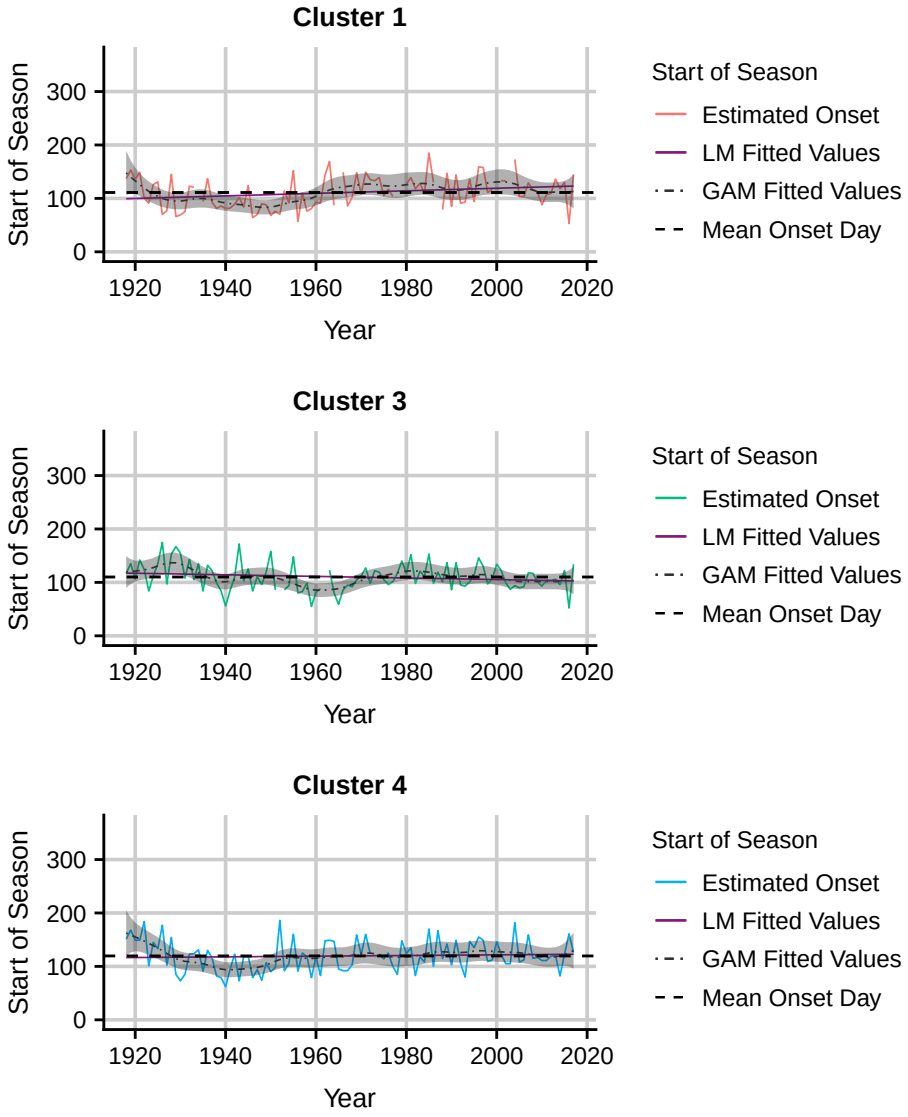


Figure 30: Start of season estimated using a threshold of probability of zero rainfall for all years (1918 to 2017). Modelling is performed on cluster daily rainfall data. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 18: Summary of the linear model fitted to the estimated onset dates using a threshold from the probability of zero rainfall model. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	Years Met
South Mediterranean	111	29.4	0.24	0.10	0.02	93
Central Mediterranean	110	25.1	-0.15	0.09	0.09	98
West Mediterranean	120	28.7	0.06	0.10	0.53	99

The linear model for the South Mediterranean cluster shows compelling evidence of linear trend with a p-value of 0.02 (*Table 18*). According to this model, the start of season day is shifting later by 0.24 days per year. The non-linear model also points toward a similar conclusion. However, between 1918 and 1948, the spline shows there is an earlier onset day to the rainfall season. This is supported by the mean onset day passing through the bottom boundary of the confidence interval around 1919. The confidence interval then shifts down until it is completely below the mean onset day between 1938 and 1956. At around 1948, the onset day starts to shift later in the year. This change is supported by the confidence interval when it is almost completely above the mean onset day (1972, 1983 and 2001) (*Figure 30, top*).

There also appears to be some evidence of linear trend for the Central Mediterranean cluster with a p-value of 0.09 (*Table 18*). According to this model, the start of season day is shifting earlier by 0.15 days per year. Looking at the non-linear trend, the GAM 95% confidence interval is completely above the mean onset day between 1925 and 1931. The confidence interval then shifts down until it is completely below the mean onset day between 1956 and 1967. This also highlights an earlier onset in the rainfall season up until 1961. The start of season remains fairly stable after that with the mean onset day falling within the confidence interval (*Figure 30, middle*).

There is no evidence of linear trend for the West Mediterranean cluster (*Table 18*). Looking at the non-linear trend, the confidence interval shifts completely above the mean onset day between 1919 and 1923 and then it shifts completely below the mean onset day between 1936 and 1949. This points to an earlier onset day until 1941. The onset day then becomes later until 1950 and thereafter, becomes very stable (*Figure 30, bottom*).

4.5.7 Start of Season Using the Gradient of Probability of Zero Rainfall

The gradient of the probability of zero rainfall models formed on the *cluster daily average rainfall* data is analysed to estimated the start of season for all years (1918 to 2017). The search is limited between the 50th and the 240th day of the year so as to exclude outlying events.

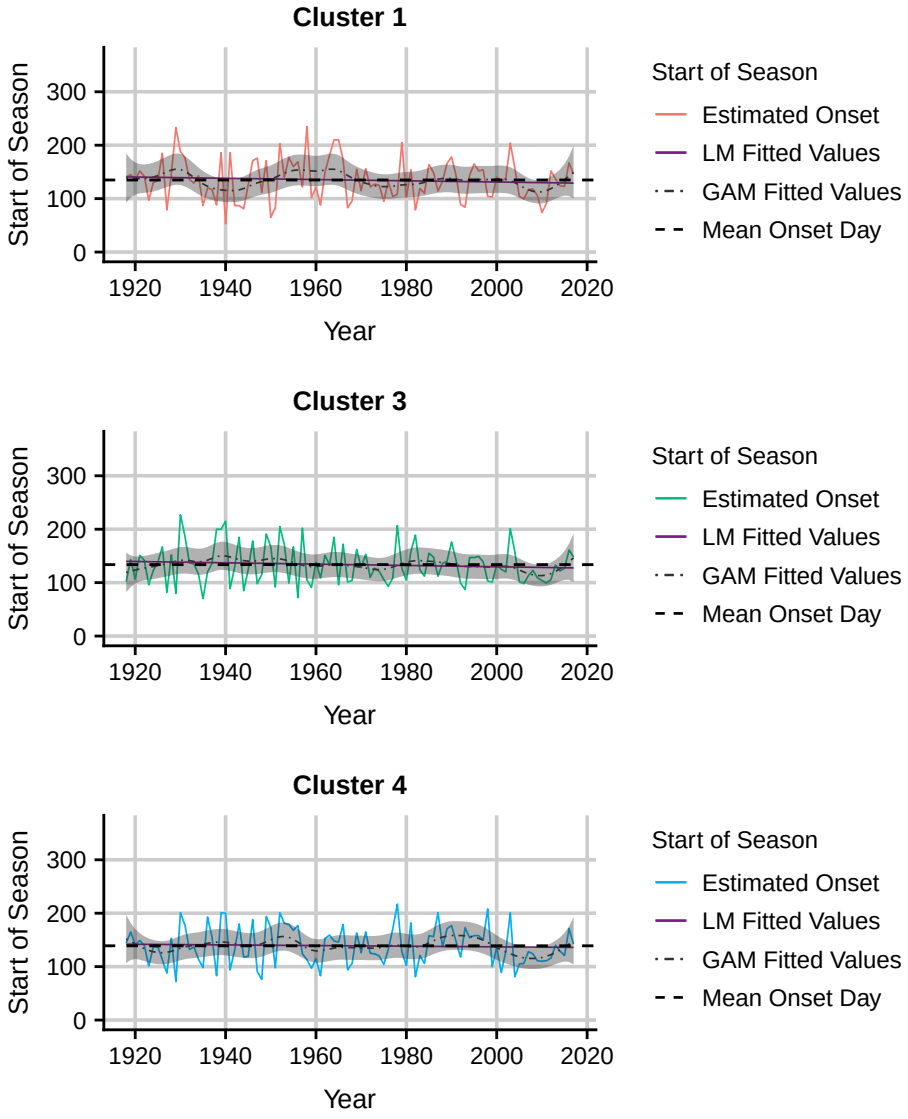


Figure 31: Start of season estimated based on the gradient of the probability of zero rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 19: Summary of the linear model fitted to the estimated onset dates using the gradient of probability of zero model. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	135	38.2	-0.11	0.13	0.38	100
Central Mediterranean	134	35.5	-0.12	0.12	0.32	100
West Mediterranean	139	35.4	-0.06	0.12	0.63	100

Using this method, all the Mediterranean clusters show no evidence of linear trend (*Table 19*).

For the South Mediterranean cluster, there is some non-linear trend. Around 1942, the confidence interval has shifted downward resulting in the mean onset day passing near the top boundary. The confidence interval is completely below the mean onset day between 2009 and 2010 and on the top of the boundary around 1954. This points to a slightly earlier start to the rainfall season up until 2009. After 2010, there is suddenly a rapid decline to the onset of the rainfall season (*Figure 31, top*).

The GAM 95% confidence interval for the Central Mediterranean cluster always contains the mean onset day except between 2009/10 where it drops completely below the mean onset day (*Figure 31, middle*).

The GAM 95% confidence interval for the West Mediterranean cluster is completely below the mean onset day between 2005 and 2010 (*Figure 31, bottom*).

4.5.8 Comparison of Different Approaches to Estimate Start of Season

The linear and non-linear trends are compared for each cluster for all of the methods used to estimate the start of season. For the linear trend, the calculated p-values (*Tables 14 - 19*) and confidence intervals of the gradient of the linear model (*Table 20*) are compared and for the non-linear trends, the years where the mean onset day does not fall within the GAM 95% confidence interval are compared for each cluster (*Tables 21 - 24*).

Abbreviations of Methods:

- 'Cum. Thr.' = 'Cumulative Threshold' (chapter 4.5.1)
- 'Stern Def.' = 'Stern Definition' (chapter 4.5.2)
- 'Mod. Def.' = 'Modified Definition' (chapter 4.5.3)
- 'Thr. Mean' = 'Threshold from Mean Daily Rainfall' (chapter 4.5.4)
- 'Thr. Pr(0)' = 'Threshold from Probability of Zero' (chapter 4.5.5)
- 'Grad. Mean' = 'Gradient from Mean Daily Rainfall' (chapter 4.5.6)
- 'Grad. Pr(0)' = 'Gradient from Probability of Zero' (chapter 4.5.7)

Table 20: Comparing the 95% confidence intervals for the gradient of the linear models for the start of season for the winter rainfall clusters (Mediterranean). Confidence intervals in bold represent confidence intervals showing evidence of change.

	Cum. Thr.		Stern Def.		Mod. Def.		Thr. Mean		Thr. Pr(0)		Grad. Mean		Grad. Pr(0)	
	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL
South Mediterranean	-0.23	0.16	-0.19	0.44	-0.23	0.10	-0.37	0.01	**0.04**	**0.44**	-0.23	0.15	-0.38	0.15
Central Mediterranean	-0.16	0.19	-0.40	0.20	-0.24	0.08	-0.24	0.12	-0.32	0.02	-0.31	0.05	-0.37	0.12
West Mediterranean	-0.26	0.15	-0.47	0.31	-0.35	0.06	**-0.54**	**-0.03**	-0.14	0.26	-0.24	0.16	-0.30	0.18

Three of the confidence intervals appear to show change (*Table 20*). This is for the South Mediterranean cluster - threshold from the GAM mean daily rainfall (chapter 4.5.4) and threshold from probability of zero rainfall (chapter 4.5.6) and then the West Mediterranean cluster - threshold from the GAM mean daily rainfall (chapter 4.5.4). These methods also all have p-values that show evidence of change.

However, the gradients showing evidence of change for the South Mediterranean cluster contradict one another (*Table 20*). The other confidence intervals for cluster 1 seem to have zero near the middle of the interval making it difficult to conclude any evidence of linear change.

Most of the confidence intervals for the Central Mediterranean and West Mediterranean clusters are on the negative side (zero falls near the positive side of the boundary) (*Table 20*). This does indicate a shift to an earlier start in season for both of these clusters.

Table 21: South Mediterranean Cluster - Years where the GAM 95% confidence interval does not contain the mean onset day. Years in italics represent the mean onset day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean onset day is completely below the confidence interval indicating an earlier start to the season relative to the mean. 'Later' refers to years where the mean onset day is completely above the confidence interval indicating a later start to the season relative to the mean.

	1918 to 1939	1940 to 1970		1971 to 1995		1996 +	
	Later	Earlier	Later	Earlier	Later	Earlier	Later
Cum. Thr.		1955 and 1959	*1971*	1980 to 1983			
Mod. Def.	1918 to 1926	1940 to 1946					
Thr. Mean				1989/90			
Grad. Mean		1947		*1989*			
Thr. Pr(0)	1919	1938 to 1956			*1972, 1983*		*2001*
Grad. Pr(0)		*1942*				2009/10	

Around the mid to late 1940s, five of the methods do suggest that some change did occur. This change is in the form of an earlier start to the season, with a shift back to a later start thereafter. Otherwise, there are not enough cases where multiple methods suggest the same change and so no other conclusion can be made (*Table 21*).

Table 22: Central Mediterranean Cluster - Years where the GAM 95% confidence interval does not contain the mean onset day. Years in italics represent the mean onset day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean onset day is completely below the confidence interval indicating an earlier start to the season relative to the mean. 'Later' refers to years where the mean onset day is completely above the confidence interval indicating a later start to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +	
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier	Later
Cum. Thr.			1957 to 1958		1981 to 1985			
Mod. Def.			1959 to 1966			*1975*	*1984*	
Thr. Mean			1947 to 1952		1986 to 1994	*1971*		
Grad. Mean			*1949*		*1987*			2004
Thr. Pr(0)		1925 to 1931	1956 to 1967					
Grad.(0)							2009/10	

Five of the methods indicate change between 1947 to 1967 with the mean onset day being below the confidence interval (*Table 22*). This suggests that there was some change with the season starting earlier and then shifting back to start later. Three of the methods suggest there was a later start around the mid to late 1980s where the start of season shifted earlier in the year.

Table 23: West Mediterranean Cluster - Years where the GAM 95% confidence interval does not contain the mean onset day. Years in italics represent the mean onset day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean onset day is completely below the confidence interval indicating an earlier start to the season relative to the mean. 'Later' refers to years where the mean onset day is completely above the confidence interval indicating a later start to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +	
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier	Later
Cum. Thr.					1979 to 1984			
Mod. Def.					*1980*		*2004*	
Thr. Mean		*1937*			*1988*		*2006*	
Grad. Pr(0)			*1965*					
Thr. Pr(0)	1936 to 1949	1919 to 1923						
Grad. Pr(0)							2005 to 2010	

There does not appear to be much change at all for the West Mediterranean cluster (*Table 23*). There is some evidence for an earlier onset date around the early 1980s. Otherwise, there are no overlapping consistent

results across the different methods and so no conclusion can be made other than the onset day has not changed over time other than the expected natural variation.

4.6 End of Season Plots

The plots in this section are all for the estimated end of season for the winter rainfall clusters (Mediterranean region). All the methods have made use of the *cluster daily average rainfall* data (*Table 2*).

4.6.1 End of Season Using a Threshold Value on Annual Cumulative Data

Different thresholds are found based on the cumulative daily rainfall anomaly (*Figure 22*). The thresholds for the South Mediterranean, Central Mediterranean and the West Mediterranean clusters are 382.0 mm, 791.1 mm and 204.3 mm, respectively (*Table 13*). There is no restriction put on the time period for when the start of season can occur within the year. However, on inspection of the estimated end of season dates, many years did not capture an end of season and so all of the threshold have been reduced based on approximately 75% of the average cumulative rainfall for the cluster over all years (1918 to 2017) (*Figure 21*). The reduced thresholds for the South Mediterranean, Central Mediterranean and the West Mediterranean clusters are 340 mm, 650 mm and 175 mm, respectively.

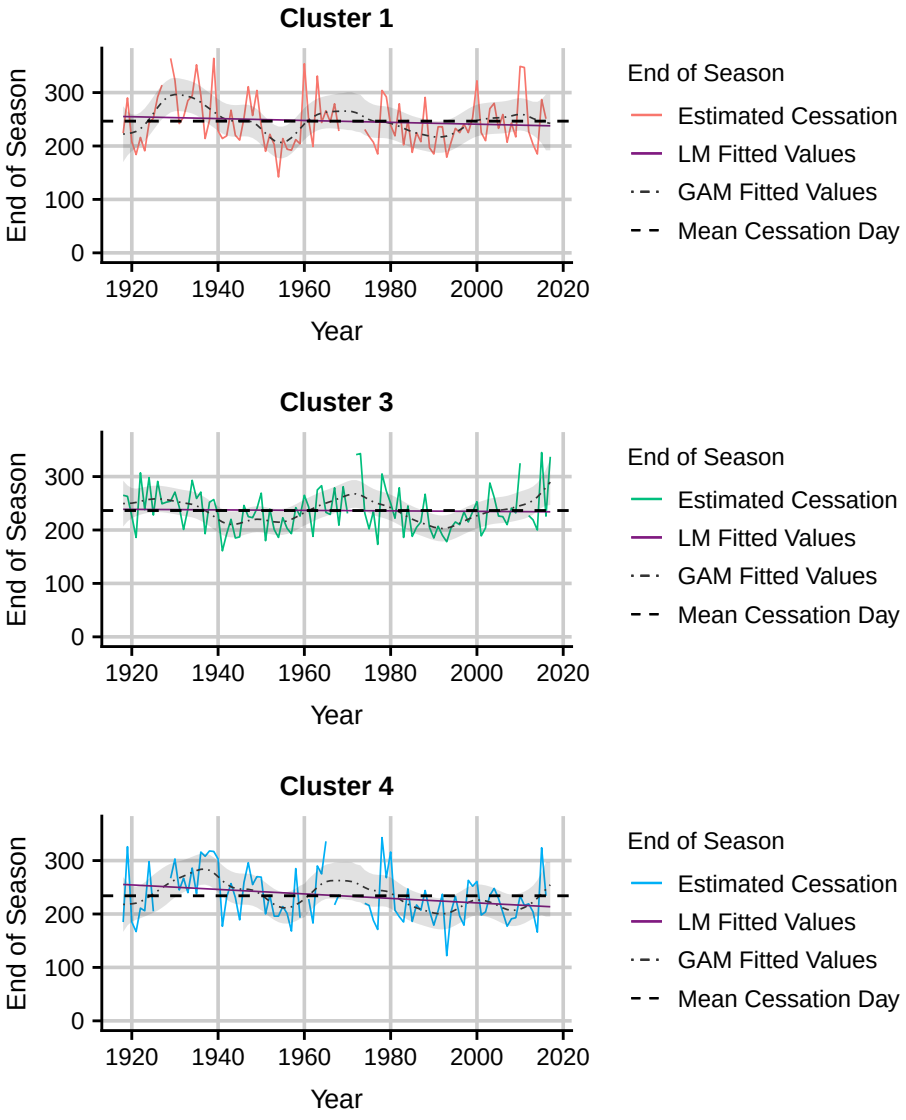


Figure 32: End of season estimated based on a threshold of cumulative data (1918 to 2017). The thresholds for clusters 1,3 and 4 are 340 mm, 650 mm and 175 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 24: Summary of the linear model fitted to the estimated cessation dates using a cumulative threshold. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	247	48.6	-0.18	0.17	0.31	95
Central Mediterranean	236	40.4	-0.05	0.14	0.75	98
West Mediterranean	234	48.0	-0.42	0.17	0.02	91

The South Mediterranean cluster shows no evidence of linear trend based on the linear model (*Table 24*) however, the GAM shows some areas of change indicating some non-linear change. The GAM 95% confidence interval is completely above the mean cessation day between 1927 and 1936. The confidence interval then shifts downward until it is completely below the mean cessation day between 1952 and 1957. Between 1990 and 1992, the confidence interval is completely above the mean cessation day (*Figure 32, top*).

The Central Mediterranean cluster also shows no linear trend (*Table 24*). There is some non-linear trend present. At around 1925, the GAM 95% confidence interval shifts upward so that the mean cessation day is right on the bottom boundary. At 1943/44, the confidence interval shifts down until it is completely below the mean cessation day. Between 1969 and 1974, the confidence interval has shifted upward such that the mean cessation day is completely below it. Between 1988 and 1995, the confidence interval has shifted downward such that the mean cessation day is completely above it. Thereafter, the confidence interval again starts to rise until is completely above the mean cessation day from 2015 onward (*Figure 32, middle*).

Only the West Mediterranean cluster shows evidence of linear trend with a p-value of 0.02 (*Table 24*). The end of season appears to be occurring 0.42 days earlier per year. Looking at the non-linear trend, the GAM confidence interval rises completely above the mean cessation day between 1931 and 1940 and then falls completely below the mean cessation day between 1988 and 1994.The confidence interval remains quite low and the mean cessation day is right on the top boundary at 2008. This highlights a later cessation to the rainfall season up until 1937. Thereafter, the cessation day appears to shift earlier in the year until roughly 1990 (*Figure 32, bottom*).

4.6.2 End of Season Based on Modified Criteria

Certain definitions are to be met for each cluster in order for the end of season to be estimated. Definitions are found in chapter 3.5.2.

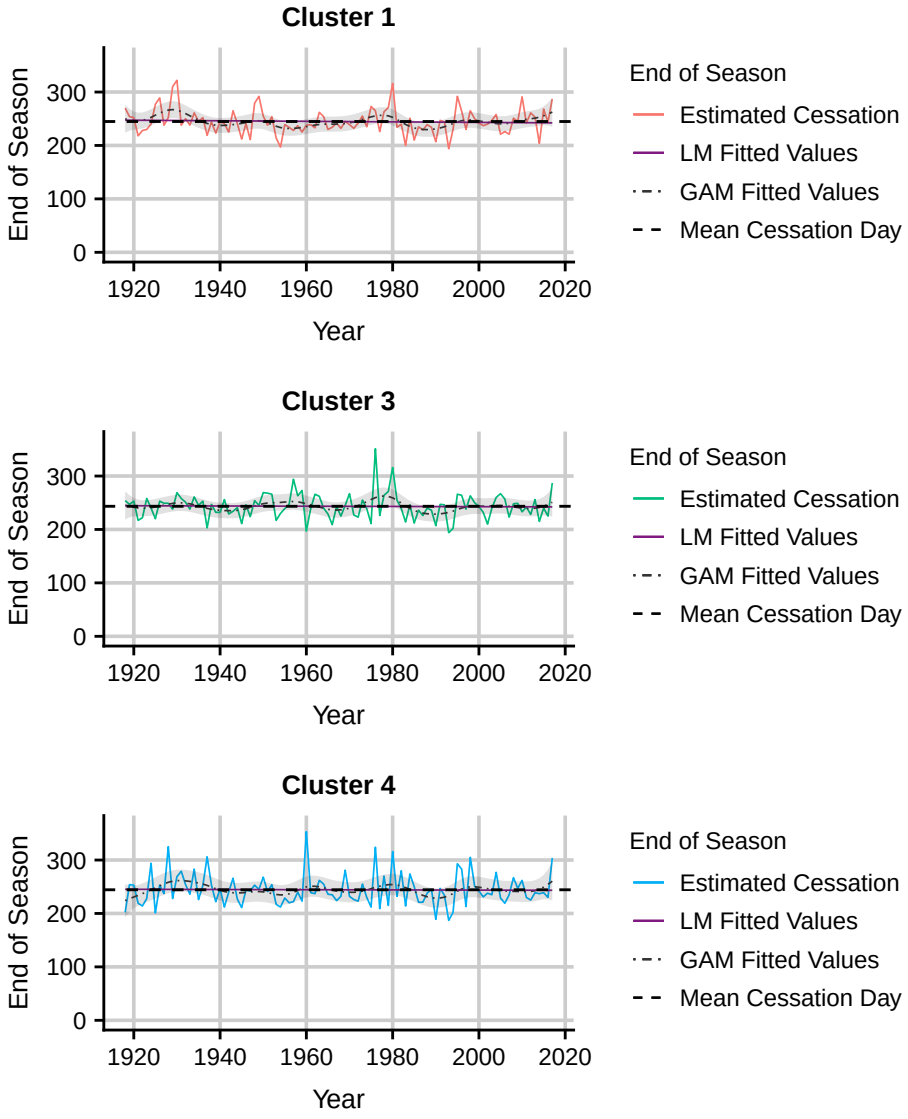


Figure 33: End of season estimated based on determined criteria (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 25: Summary of the linear model fitted to the estimated cessation dates using modified criteria. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	245	23.9	-0.05	0.08	0.55	100
Central Mediterranean	243	23.6	-0.02	0.08	0.80	100
West Mediterranean	244	30.2	-0.02	0.11	0.89	100

Based on the linear model, there is no evidence of linear trend for the Mediterranean clusters (*Table 25*).

When looking at the non-linear trend of the South Mediterranean cluster, the mean cessation day twice falls outside of the boundary of the GAM 95% confidence interval - the confidence interval rises completely above the mean cessation day between 1927 and 1931 and then shifts completely below between 1986 and 1990. Around 1956, the mean cessation day passes through the top boundary of the confidence interval. The rest of the time the cessation day is fairly stable (*Figure 33, top*).

For the non-linear trend of the Central Mediterranean cluster, there is a peak between 1977 and 1979 where the confidence interval shifts completely above the mean cessation day. The confidence interval then shifts downward until it is completely below the mean cessation day between 1989 and 1990. Otherwise, the cessation day remains constant (*Figure 33, middle*).

For the West Mediterranean cluster, even though the GAM is quite bumpy, the confidence interval always contains the mean cessation day. Around 1990, the mean cessation day passes through the top boundary of the confidence interval. Otherwise the cessation day appears stable (*Figure 33, bottom*).

4.6.3 End of Season Based on Threshold of the GAM Mean Daily Rainfall

The same thresholds are used as when computing the start of season. The search for the end of the season was limited to the 180th day of the year onward for all clusters so as to exclude outlying events. The thresholds for the South Mediterranean, Central Mediterranean and West Mediterranean clusters are 2.75 mm, 7.0 mm and 2.2 mm, respectively.

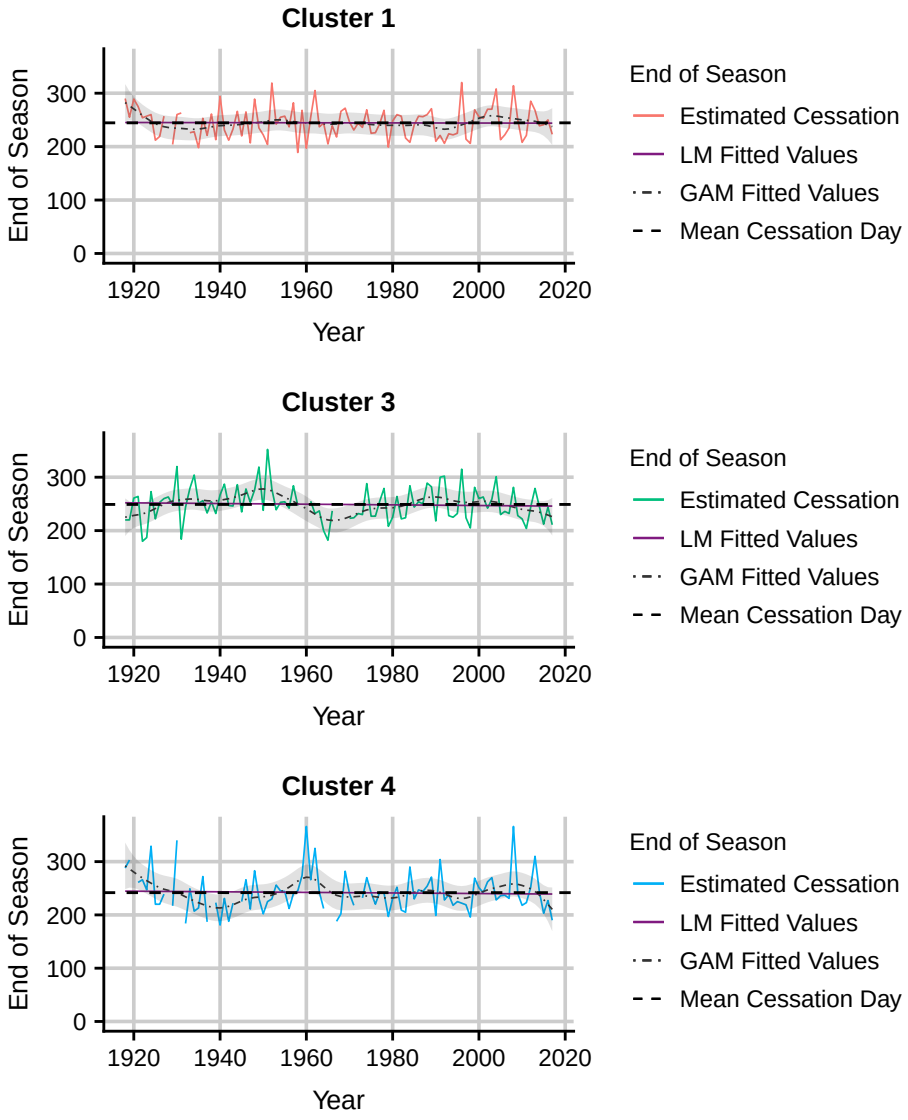


Figure 34: End of season estimated based on threshold from GAM mean daily rainfall (1918 to 2017). The thresholds for the South Mediterranean, Central Mediterranean and West Mediterranean clusters are 2.75 mm, 7 mm and 2.2 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 26: Summary of the linear model fitted to the estimated cessation dates using a threshold from the GAM mean daily rainfall. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	245	29.4	-0.01	0.10	0.91	98
Central Mediterranean	249	32.2	-0.06	0.11	0.58	96
West Mediterranean	242	37.5	-0.05	0.14	0.69	92

All of the Mediterranean clusters show no evidence of linear trend (*Table 26*).

For the South Mediterranean cluster, the non-linear trend of the GAM model supports the conclusion from the linear model. The GAM 95% confidence interval only once is above the mean cessation day from 1918 to 1920 (*Figure 34, top*).

For the non-linearity trend of the Central Mediterranean cluster. Initially, the confidence interval is low down and the mean cessation day passes through the top boundary of the confidence interval (1921). Thereafter, the confidence interval rises completely above the mean cessation day between 1946 and 1953. Between 1963 and 1970, the confidence interval shifts completely below the mean cessation day (*Figure 34, middle*).

For the West Mediterranean cluster, from 1918 to 1922, the GAM confidence interval is completely above the mean cessation day and between 1937 and 1942, the confidence interval shifts completely below the mean cessation day. Between 1959 and 1961, the confidence interval has shifted up completely above the mean cessation day (*Figure 34, bottom*).

4.6.4 End of Season Based on Threshold of Probability of Zero Rainfall

The same thresholds are used as when calculating the start of season using the probability of zero rainfall model. The search for the end of the season is limited to the 180th day of the year onward. The thresholds for the South Mediterranean, Central Mediterranean and West Mediterranean clusters are 0.5, 0.65 and 0.65, respectively.

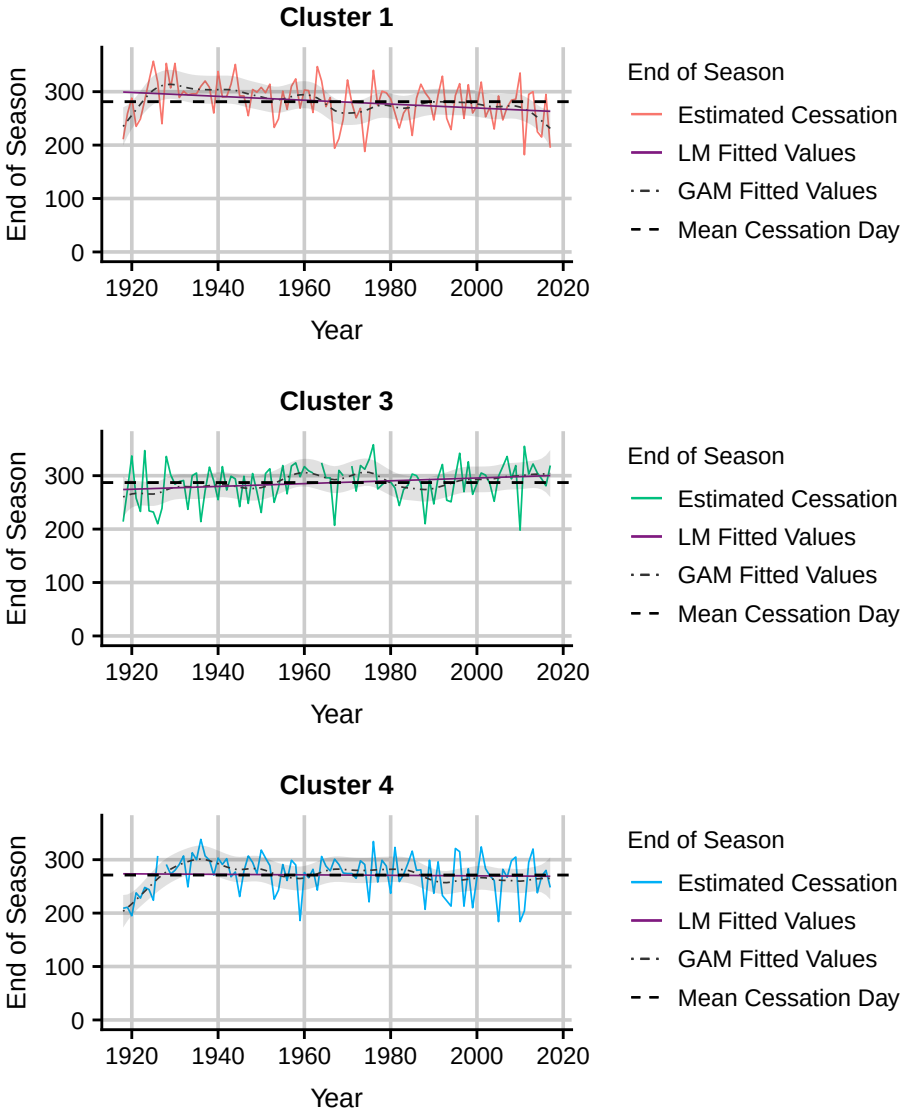


Figure 35: End of season estimated based on a threshold from probability of zero rainfall (1918 to 2017). The thresholds for the South Mediterranean, Central Mediterranean and West Mediterranean clusters are 0.5, 0.5 and 0.65, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 27: Summary of the linear model fitted to the estimated cessation dates using a threshold from the probability of zero rainfall model. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	281	38.5	-0.36	0.13	0.01	100
Central Mediterranean	287	35.0	0.26	0.12	0.03	98
West Mediterranean	271	36.2	-0.05	0.13	0.72	99

For the South Mediterranean cluster, there is strong evidence that the end of season day is shifting earlier in the year by 0.36 days per year. This is emphasized by the p-value of 0.01 (*Table 27*). This conclusion is also supported by the GAM. Initially, the GAM 95% confidence interval is completely below the mean cessation day (1918/19). The confidence interval then shifts upward until it is completely above the mean cessation day between 1927 and 1932. The confidence interval then starts to shift down until the mean cessation day passes through the top boundary around 1970 and then it passes completely below the mean cessation day from 2014 onward. This points to a slightly earlier cessation day from 1935 up until the end of the time series (*Figure 35, top*).

The Central Mediterranean cluster shows compelling evidence that the end of season is shifting 0.26 days later in the year per year (*Table 27*). This is also supported by the GAM. The GAM 95% confidence interval shifts up so that the mean cessation day passes through the top boundary around 1925. Thereafter, the confidence interval steadily rises. However, the mean cessation day remains within the boundaries. Around 1974, the mean cessation day passes near the bottom boundary. This also highlights a later cessation to the rainfall season (*Figure 35, middle*).

For the West Mediterranean cluster, there is no evidence of linear trend (*Table 27*). Looking at the non-linear trend, from 1918 to 1924,the GAM confidence interval is completely below the mean cessation day. The confidence interval then rises until it is completely above the mean between 1934 and 1938. The confidence then starts to gradually decline until around 2005. This indicates a later cessation day from 1918 to 1936, thereafter the cessation day becomes slightly earlier in the year (*Figure 35, bottom*).

4.6.5 End of Season Based on the Gradient of the GAM Mean Daily Rainfall

The gradient of the GAM mean daily rainfall is analysed to estimate the end of season. The models are formed using the *cluster daily average rainfall* for all years (1918 to 2017). The search for the end of season is limited from the 180th day since the 1st of July onward.

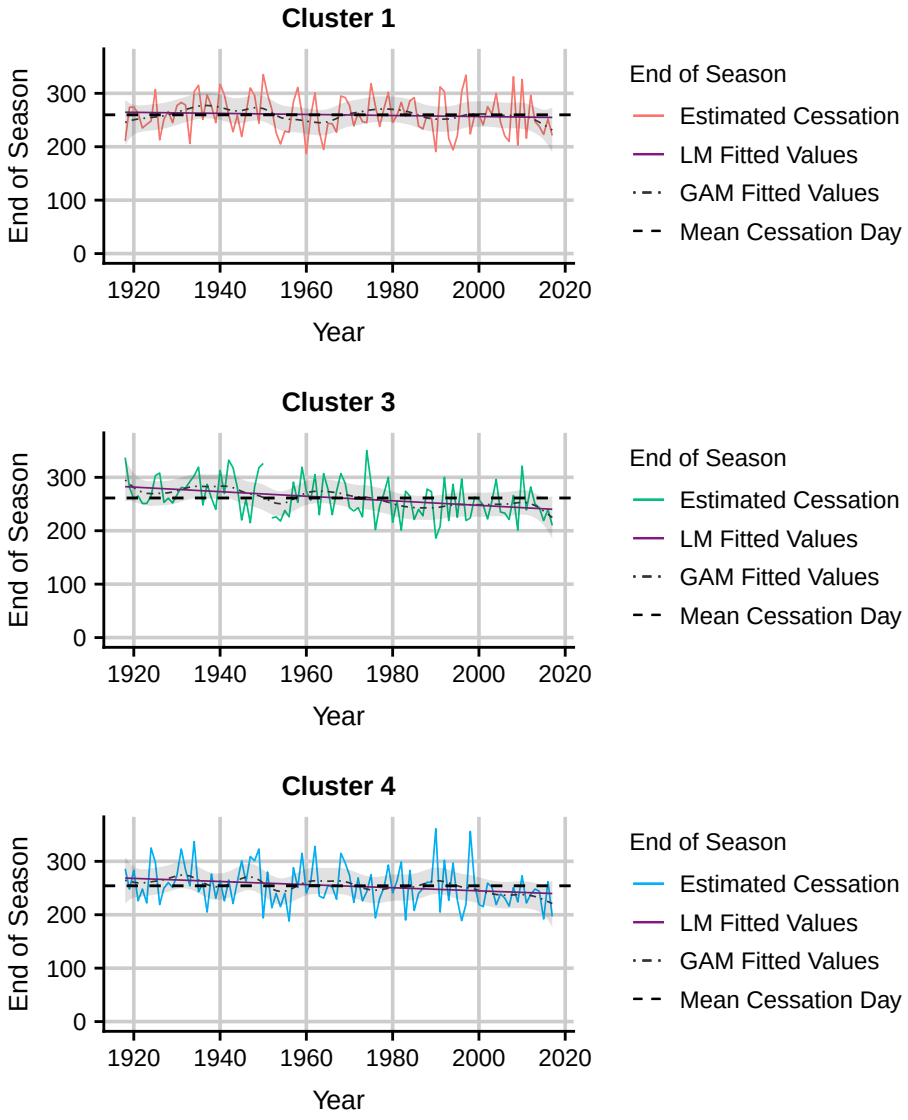


Figure 36: End of season estimated based on gradient of GAM mean daily rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 28: Results of the linear model fitted to the estimated cessation dates using the gradient of the GAM mean daily rainfall. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Standard Deviation	Gradient	Standard Error	p-value	
South Mediterranean	260	35.7	-0.10	0.12	0.43	100
Central Mediterranean	261	35.8	-0.42	0.11	0.00	99
West Mediterranean	254	37.9	-0.29	0.13	0.03	100

The South Mediterranean cluster shows no evidence of a linear trend based on its p-value (*Table 28*). Even though the GAM is quite bumpy, the bumps are fairly small and the mean onset day always fall within the GAM 95% confidence interval (*Figure 36, top*).

The Central Mediterranean cluster shows strong evidence of a decreasing linear trend based on its p-value of 0.00 (*Table 28*). The cessation day appears to be shifting 0.42 days earlier per year. Based on the non-linear trend, initially, the GAM 95% confidence interval is quite elevated and the mean cessation day passes through the bottom boundary around 1919 and then the confidence interval is completely above the mean at 1941. The confidence interval then starts to shift downward until the mean passes through the top boundary around 1987 and 2016 (*Figure 36, middle*).

The West Mediterranean cluster also shows evidence of a decreasing linear trend based on its p-value of 0.03 (*Table 28*). Even though the GAM is quite bumpy the mean cessation day is always within the boundary of the confidence interval. Around 2015, the confidence interval has shifted downward so that the mean cessation day passes near the top boundary (*Figure 36, bottom*).

4.6.6 End of Season Based on the Gradient of the Probability of Zero Rainfall

The gradient of the probability of zero rainfall model is analysed to estimate the end of season. The models are formed using the *cluster daily average rainfall* for all years (1918 to 2017). The search for the end of season is limited from the 180th day since the 1st of July onward.

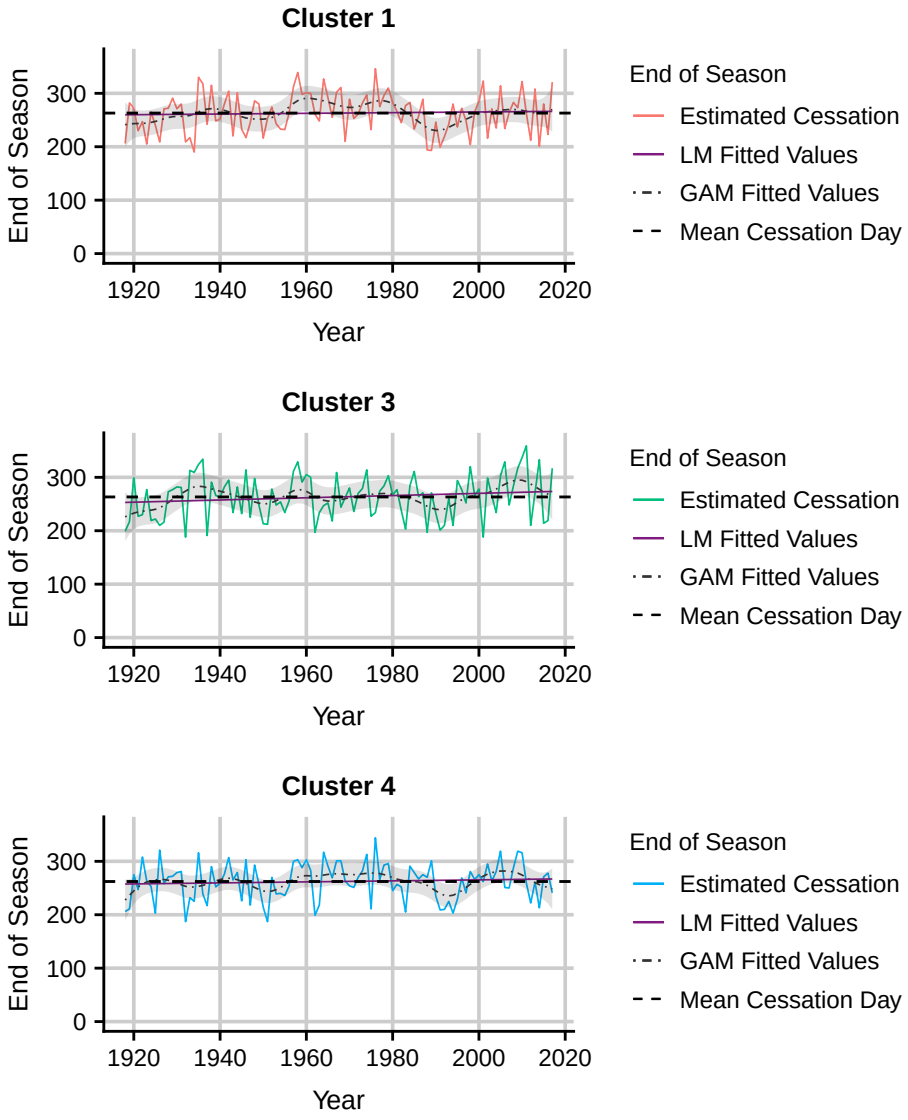


Figure 37: End of season estimated based on the gradient of probability of zero rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 29: Summary of the linear model fitted to the estimated cessation dates using the gradient of the probability of zero rainfall model. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Mediterranean	263	37.4	0.06	0.13	0.62	100
Central Mediterranean	263	40.6	0.21	0.14	0.14	100
West Mediterranean	262	34.6	0.10	0.12	0.43	100

There is no evidence of linear trend for the Mediterranean clusters (*Table 29*). Looking at the non-linearity of the South Mediterranean cluster, initially, the GAM 95% confidence interval is quite low and the mean cessation day passes near the top boundary around 1922. The confidence interval starts rising until it is completely above the mean cessation day between 1959 and 1963. Around 1977, the mean cessation day passes through the bottom boundary. Thereafter, the confidence interval starts to shift downward until it is completely below the mean cessation day between 1988 and 1993 (*Figure 37, top*).

For the Central Mediterranean cluster, initially, the GAM 95% confidence interval is below the mean cessation day from 1920 to 1922. The confidence interval does then rise but the mean cessation day remains comfortably in the confidence interval. Around 1990, the confidence interval shifts down again and the mean cessation day passes near the top boundary. The confidence interval then rises until it is completely above the mean cessation day between 2007 and 2011 (*Figure 37, middle*).

The West Mediterranean cluster initially has the confidence interval quite low and the mean cessation day passes through the top boundary (1918). Around 1951, the confidence interval has shifted downward such that the mean cessation day passes through the top boundary. Between 1991 and 1995, the confidence interval has shifted completely below the mean cessation day. Thereafter, the confidence interval rises and the mean cessation day passes through the bottom boundary (2006)(*Figure 37, bottom*).

4.6.7 Comparison of Different Approaches to Estimate the End of Season

The linear and non-linear trends are compared for each cluster for all of the methods used to determine the end of season. For the linear trend, the calculated p-values (*Tables 24 - 29*) and confidence intervals of the gradient of the linear model (*Table 30*) are compared and for the non-linear trends, the years where the mean cessation day does not fall within the GAM 95% confidence interval are compared for each cluster (*Tables 31 - 33*).

Abbreviations of Methods:

- 'Cum. Thr.' = 'Cumulative Threshold' (chapter 4.6.1)
- 'Mod. Def.' = 'Modified Definition' (chapter 4.6.2)
- 'Thr. Mean' = 'Threshold from Mean Daily Rainfall' (chapter 4.6.3)
- 'Thr. Pr(0)' = 'Threshold from Probability of Zero' (chapter 4.6.4)
- 'Grad. Mean' = 'Gradient from Mean Daily Rainfall' (chapter 4.6.5)
- 'Grad. Pr(0)' = 'Gradient from Probability of Zero' (chapter 4.6.6)

Table 30: Comparing the 95% confidence intervals for the gradient of the linear models for the end of season for the winter, Mediterranean rainfall clusters. Confidence intervals in bold represent confidence intervals showing evidence of change.

	Cum. Thr.		Mod. Def.		Thr. Mean		Grad. Mean		Thr. Pr(0)		Grad. Pr(0)	
	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL
South Mediterranean	-0.52	0.17	-0.22	0.11	-0.22	0.19	-0.34	0.15	**-0.62**	**-0.11**	-0.19	0.32
Central Mediterranean	-0.33	0.24	-0.18	0.14	-0.28	0.16	**-0.66**	**-0.19**	**0.03**	**0.50**	-0.07	0.48
West Mediterranean	**-0.76**	**-0.08**	-0.23	0.19	-0.33	0.22	**-0.55**	**-0.04**	-0.30	0.21	-0.14	0.33

The confidence intervals showing evidence of change match all the methods that resulted in p-values showing evidence of change (*Table 30*). However, only one of the methods for the South Mediterranean cluster shows evidence of linear trend. The other methods for the South Mediterranean cluster do also appear slightly on the negative side but there is not enough evidence to conclude that there is any linear change.

For the Central Mediterranean Cluster, the two methods indicating linear change contradict one another. The other methods also contradict one another and there is no evidence for linear change. For the West Mediterranean cluster, the two methods suggesting linear change both indicate a negative trend - an earlier end to the season. This is also supported by most of the other methods showing confidence intervals on the negative side (*Table 30*). Therefore, there is some evidence to suggest a linear change to an earlier end to the season for the West Mediterranean cluster.

Table 31: South Mediterranean Cluster - Years where the GAM 95% confidence interval does not contain the mean cessation day. Years in italics represent the mean cessation day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean cessation day is completely below the confidence interval indicating an earlier end to the season relative to the mean. 'Later' refers to years where the mean cessation day is completely above the confidence interval indicating a later end to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier
Cum. Thr.		1927 to 1936	1952 to 1957			1990 to 1992	
Mod. Def.		1927 to 1931	*1956*		1986 to 1990		
Thr. Mean		1918 to 1920					
Thr. Pr(0)	1918/1919	1927 to 1932			*1970*		2014+
Grad. Mean							
Grad. Pr(0)	*1922*			1959 to 1963	1988 to 1993	*1977*	

Three of the methods result in the mean onset day being above the GAM 95% confidence interval during the late 1920s/early 1930s (*Table 31*). This indicates that there was a shift to a later end to the season around then with the cessation day returning to the mean thereafter.. The confidence interval did leave the mean cessation day on multiple other occasions but this was not echoed across multiple methods.

Table 32: Central Mediterranean Cluster - Years where the GAM 95% confidence interval does not contain the mean cessation day. Years in italics represent the mean cessation day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean cessation day is completely below the confidence interval indicating an earlier end to the season relative to the mean. 'Later' refers to years where the mean cessation day is completely above the confidence interval indicating a later end to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +	
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier	Later
Cum. Thr.		1927 to 1936	1952 to 1957			1990 to 1992		
Mod. Def.					1989 to 1990	1977 to 1979		
Thr. Mean	*1921*		1963 to 1970	1946 to 1953				
Thr. Pr(0)	*1925*					*1974*		
Grad. Mean		*1919*		1941	*1987*		*2016*	
Grad. Pr(0)	1920 to 1922				*1990*			2007 and 2011

Analyzing the non-linear trend of the Central Mediterranean cluster, multiple times the confidence interval left the mean cessation day, however this not supported from method to method (*Table 32*). This makes it very difficult to conclude that there has been any non-linear change other than natural variation.

Table 33: West Mediterranean cluster - Years where the GAM 95% confidence interval does not contain the mean cessation day. Years in italics represent the mean cessation day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean cessation day is completely below the confidence interval indicating an earlier end to the season relative to the mean. 'Later' refers to years where the mean cessation day is completely above the confidence interval indicating a later end to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +	
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier	
Cum. Thr.		1931 to 1940			1988 to 1994		*2008*	
Mod. Def.							*1990*	
Thr. Mean	1937 to 1942	1918 to 1922		1959 to 1961				
Thr. Pr(0)	1918 to 1924	1934 to 1938						
Grad. Mean							*2015*	
Grad. Pr(0)	*1918*		*1951*				1991 and 1995	*2006*

The West Mediterranean cluster produces a similar result to the non-linear changes of the Central Mediterranean cluster. Again, the confidence interval often leaves the mean cessation day but this is not supported from method to method (*Table 33*).

4.7 Start of Season Plots for SUMMER RAINFALL Clusters

Start of season for the summer rainfall clusters, the South Coast and Karoo clusters, are estimated. In order to estimate the start of season for the summer rainfall regions, days between 1 July n to 30 June (n+1) are looked through. Thus, the year 1919 refers to: 1 July 1918 to 31 June 1919. The *cluster daily average rainfall* data (*Table 2*) is used for all of these methods.

4.7.1 Start of Season Using a Threshold Value based on Cumulative Annual Rainfall

Different thresholds are found based on the cumulative daily rainfall anomaly (*Figure 22*). The thresholds for the South Coast (2) and Karoo (5) clusters are 234.1 mm, 115.1 mm, respectively (*Table 13*). There is no limit on the search for the start of season for all of the years (1918 to 2017).

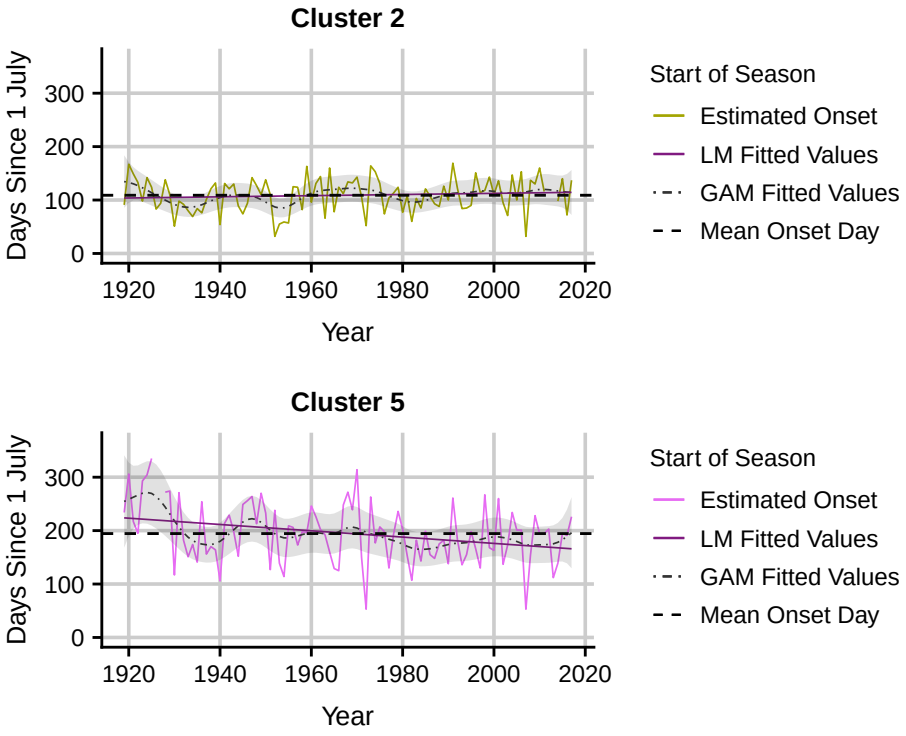


Figure 38: Start of season estimated based on threshold of cumulative data (1918 to 2017). The thresholds for the South Coast and Karoo clusters are 234.1 mm and 115.1 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 34: Summary of the linear model fitted to the estimated onset dates using a cumulative threshold. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	109	31.7	0.11	0.11	0.35	97
The Karoo	194	55.5	-0.59	0.19	0.00	97

The South Coast cluster (2) shows no evidence of linear trend based on the linear model and its p-value (*Table 34*). The South Coast cluster does appear to have had some changes in onset to the rainfall season.

The 95% GAM confidence interval is completely below the mean onset day between 1931 and 1935 as well as between 1952 to 1955 (*Figure 38, top*).

The Karoo cluster (5) does show evidence of linear trend with a small p-value of 0.00 (*Table 34*). The onset day appears to be shifting 0.59 days earlier in the year. The GAM also supports this linear trend. The GAM 95% confidence interval falls completely above the mean onset day between 1921 and 1927. Thereafter, the mean onset day remains completely within the confidence interval, however the confidence interval keeps shifting downward such that the mean onset day is near the top boundary around 1984. This plot does have a really wide range in onset days (between 53 and 335 days since 1 July). The Karoo cluster does cover a region that does have aseasonal rainfall and this is why there might be such a big range (*Figure 38, bottom*).

4.7.2 Start of Season Based on Modified Criteria

Certain criteria are to be met for each cluster in order for the start of season to be estimated. Criteria are found in chapter 3.5.2.

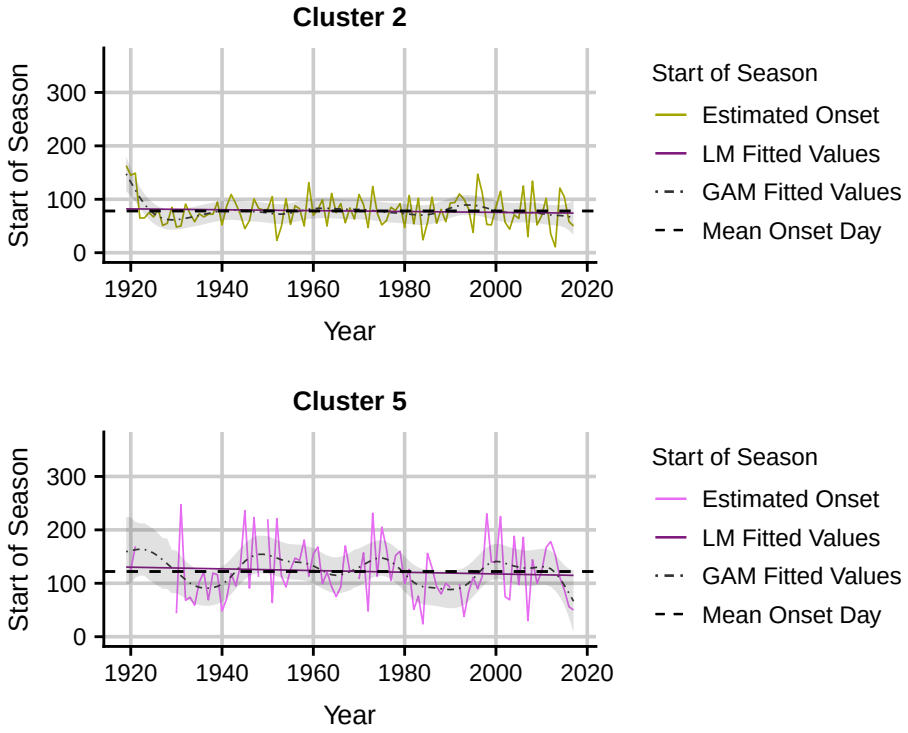


Figure 39: Start of season estimated based on modified criteria (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 35: Summary of the linear model fitted to estimated onset dates using modified criteria. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	78	29.4	-0.08	0.10	0.44	99
The Karoo	122	52.5	-0.16	0.21	0.44	89

Neither clusters show any evidence of linear trend based on their p-values (*Table 35*). For the non-linear trend of the South Coast cluster, the GAM 95% confidence interval is completely above the mean onset day between 1919 and 1922. The confidence interval remains fairly constant after this (*Figure 39, top*).

For the Karoo cluster, the confidence interval shifts down around 1937 resulting in the mean onset day passing near the top boundary. The confidence interval then shifts upward until around 1948 where the mean onset day passes near the bottom boundary. Then around 1985 to 1992, the confidence interval has shifted downward and the mean onset day is near the top boundary. Lastly, the confidence interval shifts down at 2017 so that the mean onset day passes through the top boundary (*Figure 39, bottom*).

4.7.3 Start of Season Based on Threshold from GAM Mean Daily Rainfall

A threshold of 7 mm is chosen for the South Coast cluster and a threshold of 3.1 mm is chosen for the Karoo cluster. These values are chosen based on the average, smoothed seasonal profile (*Figure 21*). For the South Coast cluster, the search is limited from the 1st to the 265th day since 1 July. The search for the Karoo cluster is limited from the 50th to the 300th day since 1 July. This is due to the strong peak of rainfall that appears later in the year.

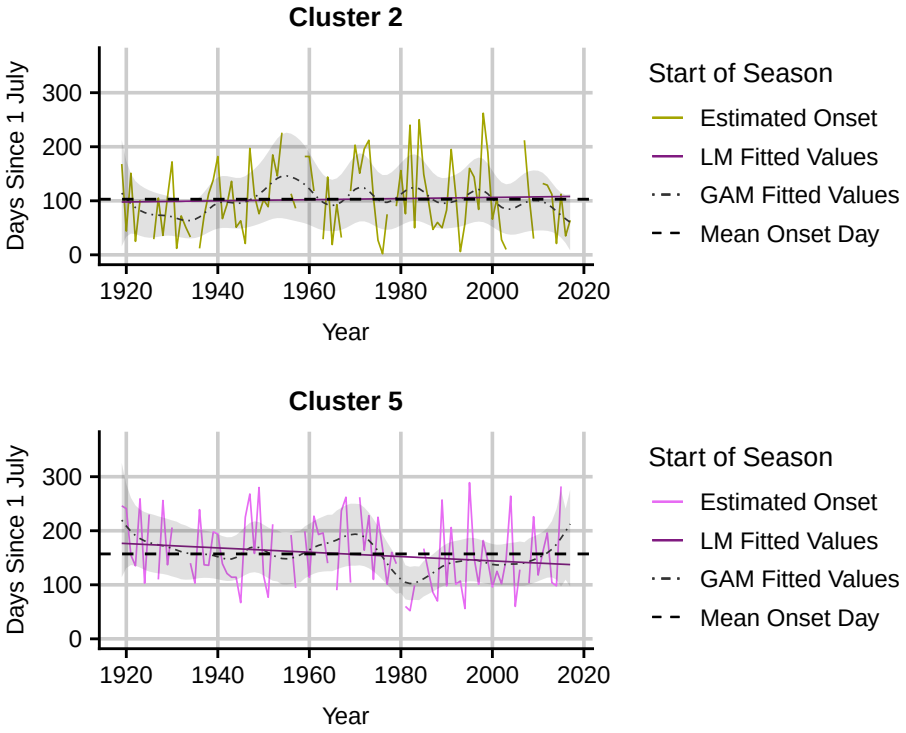


Figure 40: Start of season estimated based on threshold of GAM mean rainfall (1918 to 2017). Thresholds for clusters 2 and 5 are 7 mm and 3.1 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 36: Summary of the linear model fitted to the estimated onset dates using a threshold from the GAM mean daily rainfall. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	103	64.3	0.10	0.24	0.67	88
The Karoo	157	63.7	-0.40	0.24	0.10	86

The South Coast cluster shows no evidence of linear trend (*Table 36*). The mean onset day mostly falls within the confidence interval. Between 1931 and 1934, the GAM confidence interval shifts completely below the mean onset day. Thereafter, the mean onset day then falls comfortably within the confidence interval. This is due to the large standard deviation of the start of season days which results in a wide confidence interval (*Figure 40, top*).

The Karoo cluster shows marginal evidence of linear trend based on its p-value of 0.10 (*Table 36*). The start of season day appears to be shifting 0.40 days earlier per year. Looking at the non-linear changes, the mean onset day mostly falls within the confidence interval. Between 1979 and 1987, the confidence interval

shifts completely below the mean onset day which points to an earlier start to the season up until 1982. Thereafter, the start of season shifts back later in the year. Similar to the South Coast cluster, the Karoo cluster also has a high standard deviation resulting in a wide confidence interval (*Figure 40, bottom*).

4.7.4 Start of Season Based on Threshold from GAM Probability of Zero Rainfall

Thresholds of 0.68 and 0.72 for the South Coast and Karoo clusters, are used respectively. These are chosen based on visual inspection of the probability of zero rainfall models. The search for the start of season is limited between the 1st and the 265th day since the 1st July for the South Coast cluster and for the Karoo cluster, the search is limited between the 50th and 300th day since 1 July.

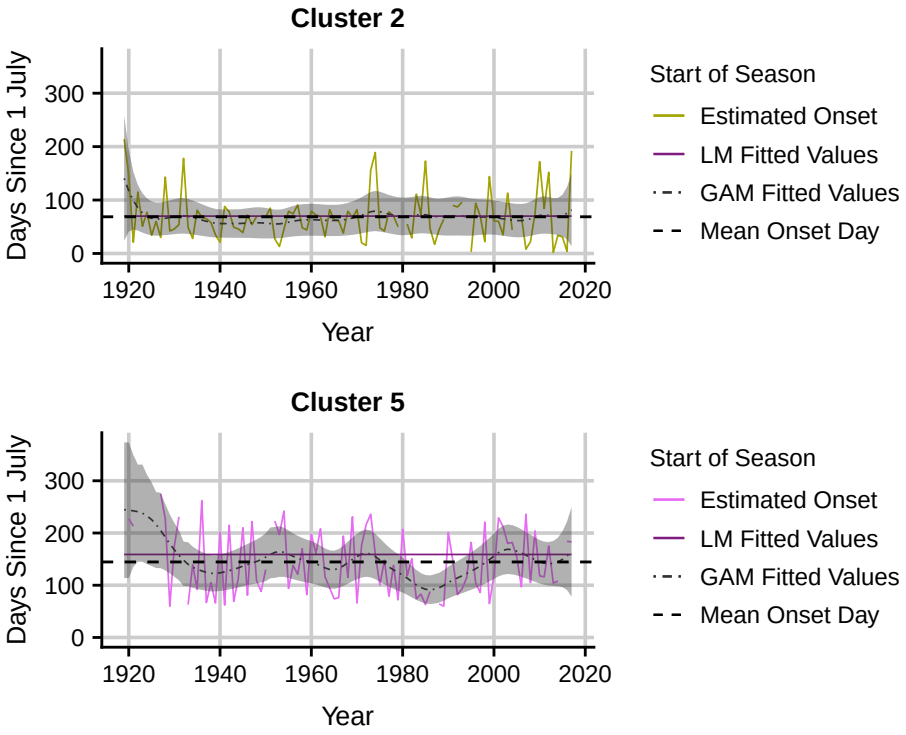


Figure 41: Start of season estimated based on threshold of probability of zero rainfall (1918 to 2017). Thresholds for clusters 2 and 5 are 0.68 and 0.72, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 37: Summary of the linear model fitted to the estimated onset dates using a threshold from the probability of zero model. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	69	44.9	-0.03	0.16	0.87	94
The Karoo	145	61.9	-0.30	0.24	0.21	90

Neither clusters show evidence of linear trend based on their p-values (*Table 37*). For the non-linearity changes of the South Coast cluster, the confidence interval is very stable and the mean onset day always falls within the boundaries (*Figure 41, top*).

For the Karoo cluster, the confidence interval is initially raised and mean onset day passes just below the bottom boundary around 1923. The confidence interval then shifts downward until it is completely below the mean onset day between 1982 and 1991. This points to an earlier start to the season up until 1986 where it then start to shift later (*Figure 41, bottom*).

4.7.5 Start of Season Based on the Gradient of the GAM Mean Daily Rainfall

The gradient of the GAM mean daily rainfall is analysed to estimate the start of season for all years (1918 to 2017). Modelling is performed on the *cluster daily average rainfall* data. The South Coast cluster is limited between the 1st and 265th day since 1 July. The Karoo cluster is limited between the 50th and the 300th day since 1 July.

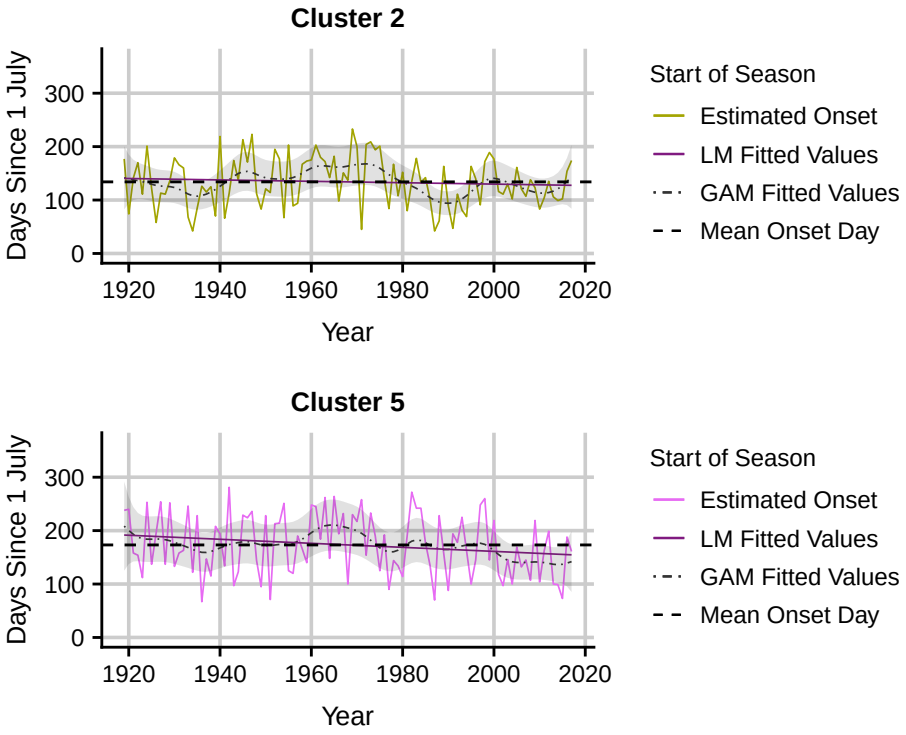


Figure 42: Start of season estimated based on the gradient of the GAM mean daily rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 38: Summary of the linear model fitted to the estimated onset dates using the gradient of the GAM mean daily rainfall. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	134	46.8	-0.13	0.17	0.43	99
The Karoo	173	57.1	-0.38	0.20	0.06	99

The South Coast cluster shows no evidence of linear trend based on its p-value from the linear model (*Table 38*). The GAM confidence interval is completely below the mean onset day between 1934/35 and again between 1986 and 1994 (*Figure 42, top*).

The Karoo cluster shows some evidence of linear trend based on its p-value of 0.06 from the linear model (*Table 38*). The start of season appears to be shifting 0.38 days earlier per year. The GAM confidence interval has a downward trend and this is highlighted when it goes completely below the mean onset day between 2005 and 2007 as well as 2011 to 2014 (*Figure 42, bottom*). This points to an earlier start in season.

4.7.6 Start of Season Based on the Gradient of the Probability of Zero Rainfall

The gradient of the probability of zero rainfall model is analysed to estimate the start of season for all years (1918 to 2017). Modelling is performed on the *cluster daily average rainfall* data. The South Coast cluster is limited between the first and 265th day since 1 July. The Karoo cluster is limited between the 50th and the 300th day since 1 July.

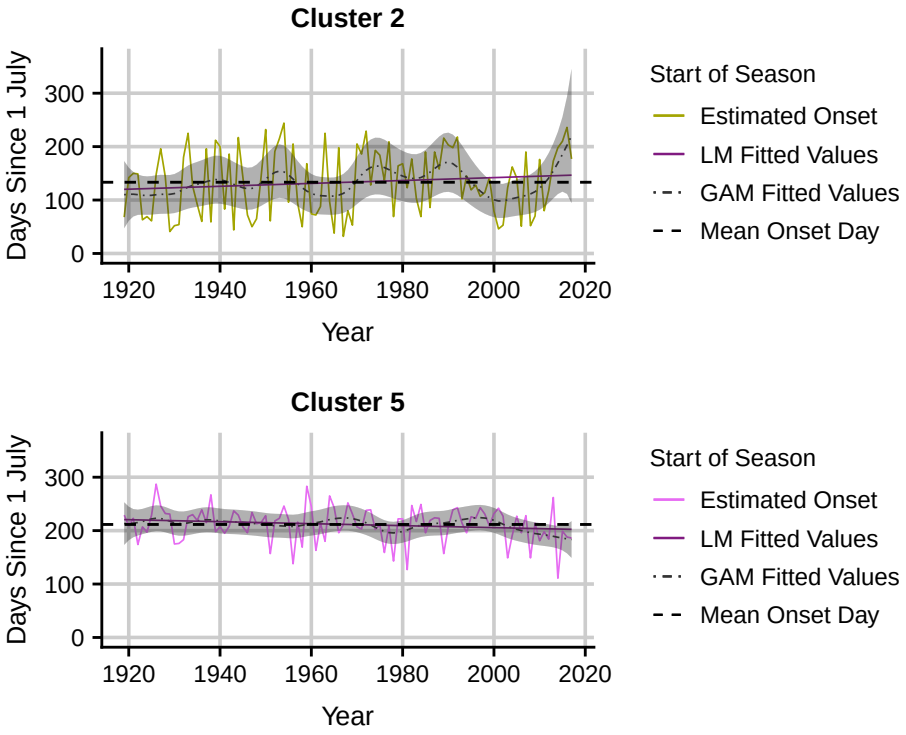


Figure 43: Start of season estimated based on the gradient of the probability of zero rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 39: Summary of the linear model fitted to the estimated onset dates using the gradient of the probability of zero rainfall model. 'n' refers to the number of years where an estimated onset date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	133	61.3	0.27	0.22	0.21	99
The Karoo	212	32.1	-0.19	0.11	0.10	99

The South Coast cluster shows no evidence of linear trend based on the linear model (*Table 39*). For its non-linear changes, the confidence interval mostly contains the mean onset day for the whole period except for between 2001 and 2003 where the confidence interval is completely below the mean onset day (*Figure 43, top*).

The Karoo cluster shows marginal evidence of linear trend based on its p-value of 0.10 (*Table 39*). Looking at its non-linear changes, the confidence interval always contains the mean onset day except between 2012 and 2016 where the confidence interval is completely below the mean onset day. Around 1978, the mean onset day passes near the top boundary of the confidence interval (*Figure 43, bottom*).

4.7.7 Comparison of Different Approaches to Estimate the Start of Season

The linear and non-linear trends are compared for each cluster for all of the methods used to determine the start of season. For the linear trend, the calculated p-values (*Tables 34 - 39*) and confidence intervals of the gradient of the linear model (*Table 40*) are compared and for the non-linear trends, the years where the mean onset day does not fall within the GAM 95% confidence interval are compared for each cluster (*Tables 41 - 43*).

Abbreviations of Methods:

- 'Cum. Thr.' = 'Cumulative Threshold' (chapter 4.7.1)
- 'Mod. Def.' = 'Modified Definition' (chapter 4.7.2)
- 'Thr. Mean' = 'Threshold from Mean Daily Rainfall' (chapter 4.7.3)
- 'Thr. Pr(0)' = 'Threshold from Probability of Zero' (chapter 4.7.4)
- 'Grad. Mean' = 'Gradient from Mean Daily Rainfall' (chapter 4.7.5)
- 'Grad. Pr(0)' = 'Gradient from Probability of Zero' (chapter 4.7.6)

Table 40: Comparing the 95% confidence intervals for the gradient of the linear models for the start of season for the summer rainfall clusters (South Coast and Karoo). Confidence intervals in bold represent confidence intervals showing evidence of change.

	Cum. Thr.		Mod. Def.		Thr. Mean		Thr. Pr(0)		Grad. Mean		Grad. Pr(0)	
	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL
South Coast	-0.12	0.34	-0.29	0.12	-0.38	0.59	-0.35	0.29	-0.46	0.20	-0.16	0.70
The Karoo	**-0.97**	**-0.21**	-0.57	0.25	*-0.88*	*0.07*	-0.78	0.17	*-0.77*	*0.02*	*-0.41*	*0.04*

The confidence intervals echo the p-values from the gradients of the linear models. Looking at the South Coast cluster, there is no evidence of linear change as all the confidence intervals contradict one another (*Table 40*).

For the Karoo cluster, all of the six methods show evidence of a negative linear trend (*Table 40*). There is compelling evidence to conclude that there is an earlier start to the rainfall season for this cluster.

Table 41: South Coast - Years where the GAM 95% confidence interval does not contain the mean onset day. Years in italics represent the mean onset day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean onset day is completely below the confidence interval indicating an earlier start to the season relative to the mean. 'Later' refers to years where the mean onset day is completely above the confidence interval indicating a later start to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier
Cum. Thr.	1931 and 1935		1952 and 1955				
Mod. Def.		1919 and 1922					
Thr. Mean	1931 and 1934						
Thr. Pr(0) thr.							
Grad. Mean	1934/35				1986 to 1994		
Grad. Pr(0)							2001 and 2003

Three of the methods show the confidence interval being below the mean onset day during the early 1930s (*Table 41*). This suggests that there a slight change to an earlier start to the season which then shifted back to the mean. There are no other overlapping years where the confidence interval left the mean onset day and so no other conclusion can be made.

Table 42: The Karoo - Years where the GAM 95% confidence interval does not contain the mean onset day. Years in italics represent the mean onset day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean onset day is completely below the confidence interval indicating an earlier start to the season relative to the mean. 'Later' refers to years where the mean onset day is completely above the confidence interval indicating a later start to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +	
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier	Later
Cum. Thr.		1921 to 1927			*1984*			
Mod. Def.	*1937*			*1948*	*1988*		*2017*	
Thr. Mean					1979 to 1987			
Thr. Pr(0)		1923			1982 to 1991			
Grad. Mean							2006, 2013	
Grad. Pr(0)					*1978*		2012 to 2016	

Around the 1980s, five of the six methods suggest that the confidence interval drops below the mean onset day. This indicates a shift to an earlier start to the season. Four of the six methods also suggest that there was another dip around 2013. This again suggest an earlier start to the season with the season then stabilizing thereafter (*Table 42*). Based on the linear model results and these non-linear results, there is compelling evidence indicating an earlier start in season for the Karoo cluster.

4.8 End of Season Plots for SUMMER RAINFALL Clusters

The end of season estimates based on the different methods. The *cluster daily average rainfall* data (*Table 2*) is used when computing all the methods for all of the years (1918 to 2017).

4.8.1 Using a Threshold Value based on Cumulative Annual Rainfall

Different thresholds are found based on the cumulative daily rainfall anomaly (*Figure 22*). The thresholds for clusters 2 and 4 are 515.2 mm, 204.2 mm, respectively (*Table 13*). On visual inspection of the estimated cessations for the Karoo cluster, a large number of the years do not exceed the threshold. Thus, the threshold for the Karoo cluster is slightly relaxed to approximately 75% of the rainfall from the average cumulative plots (*Figure 21*). The thresholds for the South Coast and Karoo clusters are 515.2 mm and 170.0 mm, respectively.

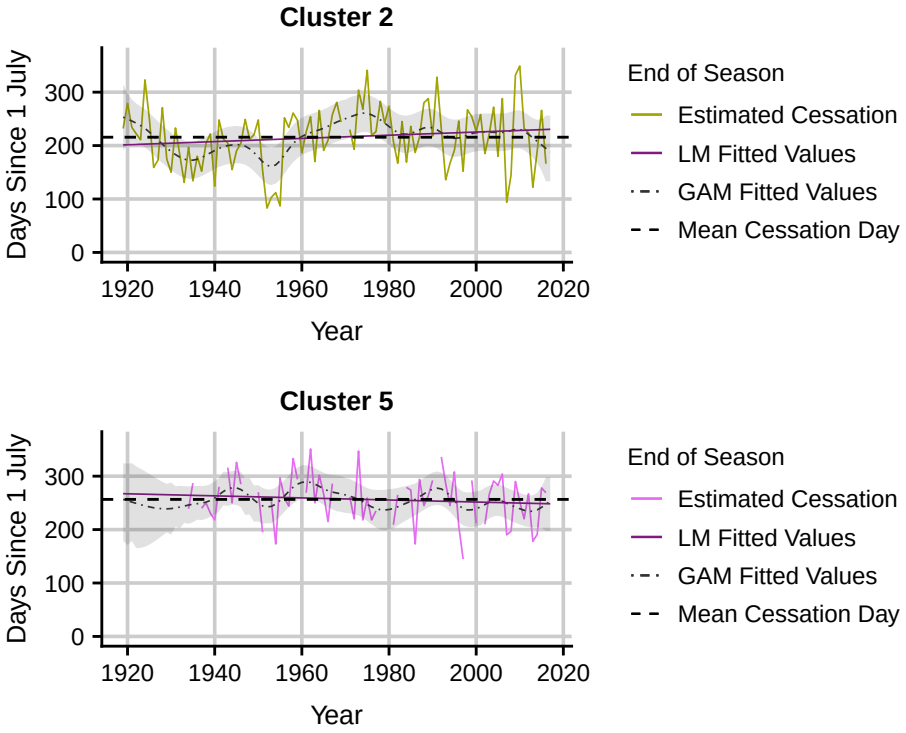


Figure 44: End of season estimated based on threshold of cumulative annual rainfall (1918 to 2017). Thresholds are 515.2 mm and 170.0 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 43: Summary of the linear model fitted to the estimated cessation dates using a cumulative threshold. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	216	56.9	0.30	0.20	0.14	97
The Karoo	245	51.2	-0.44	0.19	0.03	90

The South Coast cluster shows slight evidence of linear change based on its p-value of 0.14 (*Table 43*). There does appear to be a slight latening to the cessation of the season (0.30 days per year). The variance

is quite high which could affect the results. Based on the non-linear changes, between 1932 and 1937 the confidence interval shifts completely below the mean cessation day. It is also completely below between 1950 and 1955 but this is as a result of 3 outlying observations. The confidence interval then shifts upward until it is completely above the mean cessation day between 1972 to 1977. The confidence interval does start to shift downward thereafter but it still completely covers the mean cessation day (*Figure 44, top*).

The Karoo cluster shows marginal evidence of linear trend with a p-value of 0.03 (*Table 43*). The cessation day is starting 0.44 days earlier per year. Looking at the non-linear trend, the GAM 95% confidence interval is only once completely out of the confidence interval. Initially, the confidence interval is elevated and the mean cessation day passes through the bottom boundary around 1921. The confidence interval then keeps shifting downward until it is completely below the mean cessation day during 2010 (*Figure 44, bottom*).

4.8.2 End of Season Based on Modified Criteria

Certain criteria are to be met for each cluster in order for the end of season to be estimated. Criteria are found in chapter 3.5.2.

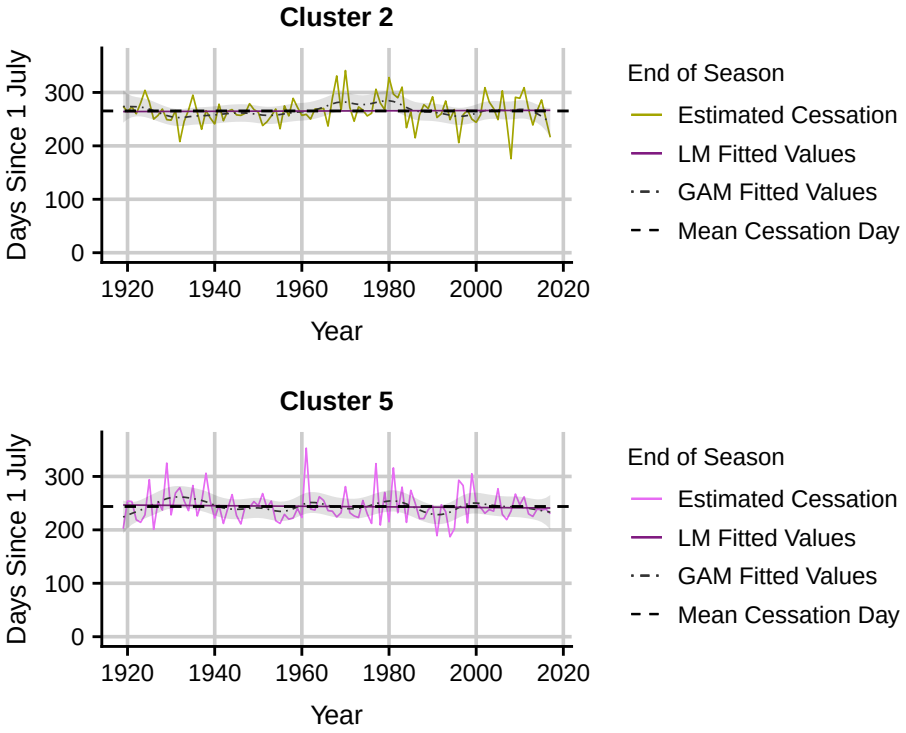


Figure 45: End of season estimated based on modified criteria (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 44: Summary of the linear model fitted to the estimated cessation dates using modified criteria. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	265	25.6	0.02	0.09	0.79	99
The Karoo	244	29.8	-0.06	0.11	0.59	99

Both the South Coast and Karoo clusters show no evidence of linear trend based on their p-values (*Table 44*). For the non-linear changes of the South Coast cluster, around 1932, the GAM 95% confidence interval shifts downward and the mean cessation day passes near the top boundary. The confidence interval then starts to rise until it is completely above the mean cessation day between 1978 and 1981 (*Figure 45, top*).

Looking at the non-linear changes of the Karoo cluster, the GAM 95% confidence interval always contains the mean cessation day. Around 1932, the mean cessation day passes through the bottom boundary of the confidence interval as it has shifted upward. The confidence interval then slowly shifts downward until 1991 where the mean cessation day passes through the top boundary (*Figure 45, bottom*).

4.8.3 End of Season Based on Threshold from GAM Mean Daily Rainfall

Thresholds of 7 mm and 2.3 mm are used for the two clusters respectively. These thresholds are chosen based on the average seasonal profiles (*Figure 21*). Both clusters are limited from the 180th day since July 1 onward to estimate the start of season. GAMs are produced based on *cluster daily average rainfall* for all years (1918 to 2017).

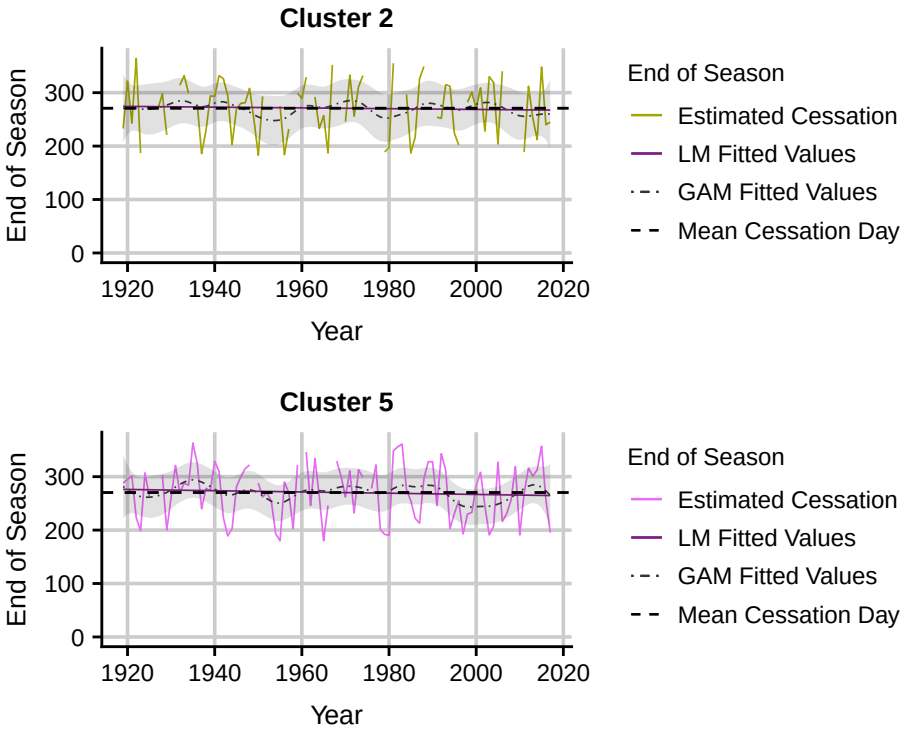


Figure 46: End of season estimated based on thresholds from GAM mean daily rainfall (1918 to 2017). The thresholds for the clusters are 7 mm and 2.3 mm, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 45: Summary of the linear model fitted to the estimated cessation dates using a threshold on the GAM mean daily rainfall. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	271	51.2	-0.07	0.20	0.73	75
The Karoo	270	50.4	-0.12	0.18	0.52	93

Many of the years do not meet the requirements for the South Coast cluster. The cluster shows no evidence of linear change (*Table 45*) and this is supported by the mean cessation day always falling comfortably in the middle of the confidence interval (*Figure 46, top*).

The Karoo cluster produces a similar result to the South Coast cluster. There is no evidence of linearity and the GAM 95% confidence interval always contains the mean cessation day (*Figure 46, bottom*).

4.8.4 End of Season Based on a Threshold from the Probability of Zero Rainfall

Thresholds of 0.7 and 0.67 are used for the two clusters respectively based on visual inspection of the probability of zero rainfall models. Both cluster are limited from the 180th day since July 1 onward.

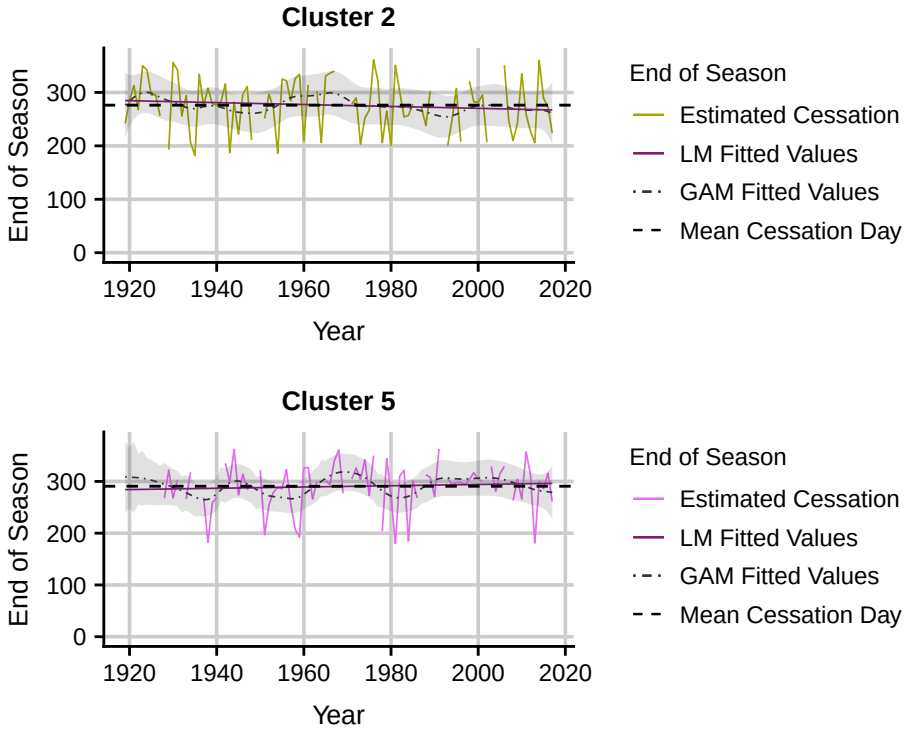


Figure 47: End of Season estimated based on a threshold from the probability of zero rainfall model (1918 to 2017). Thresholds for the South Coast and Karoo clusters are 0.7 and 0.67, respectively. The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 46: Summary of the linear model fitted to the estimated cessation dates using a threshold on the probability of zero rainfall model. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	276	48.3	-0.18	0.18	0.32	86
The Karoo	291	42.6	0.12	0.18	0.49	80

The South Coast cluster produces a very similar result to the above method results (threshold on the expected daily rainfall). There is no evidence of linear trend (*Table 46*) and the confidence interval is always comfortably covering the mean cessation day indicating no non-linear trend (*Figure 47, top*).

The Karoo cluster also shows no evidence of linear trend (*Table 46*). Looking at the non-linear trend, around 1957, the confidence interval shifts downward and the mean cessation day passes through the top boundary (*Figure 47, bottom*).

4.8.5 End of Season Based on the Gradient of the GAM Mean Daily Rainfall

The end of season is estimated based on the gradient of the GAM expected daily rainfall. These models are produced from the *cluster daily average rainfall* data for all years (1918 to 2017). The search for the end of season day is limited from the 180th day since 1st July onward.

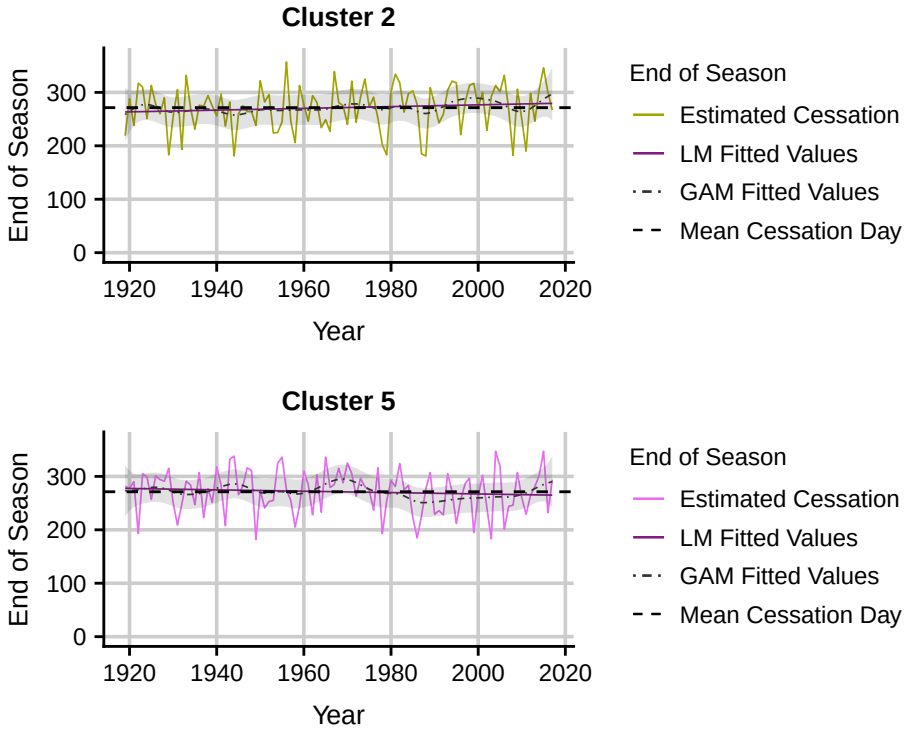


Figure 48: End of season estimated based on the gradient of the GAM mean daily rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 47: Summary of the linear model fitted to the estimated cessation dates using the gradient of the GAM mean daily rainfall. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	271	41.7	0.16	0.15	0.26	99
The Karoo	271	40.3	-0.13	0.14	0.37	99

The South Coast and Karoo clusters show no evidence of linear trend (*Table 47*). For the non-linear changes of the South Coast cluster, the GAM 95% confidence interval always contains the mean cessation day (*Figure 48, top*).

For the Karoo cluster, around 1969, the confidence interval shifts upward and the mean cessation day passes near the bottom boundary of the confidence interval. Otherwise, the confidence interval always covers the mean cessation day (*Figure 48, bottom*).

4.8.6 End of Season Based on the Gradient of Probability of Zero Rainfall

The end of season is estimated based on the gradient of the probability of zero rainfall model. These models are produced from the *cluster daily average rainfall* data for all years (1918 to 2017). The search for the end of season day is limited from the 180th day since 1st July onward.

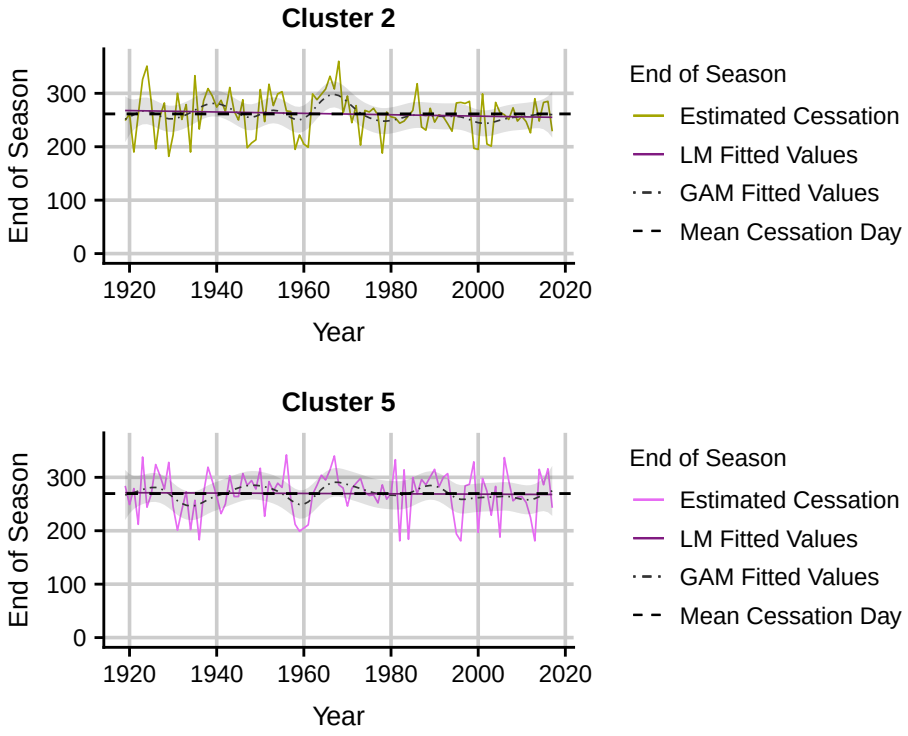


Figure 49: End of season estimated based on the gradient of the probability of zero rainfall (1918 to 2017). The linear model closely corresponds to the mean and that is why it is difficult to see. The shaded area indicates a 95% confidence interval fitted to the GAM.

Table 48: Summary of the linear model fitted to the estimated cessation dates using the gradient of the probability of zero rainfall model. 'n' refers to the number of years where an estimated cessation date is captured.

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value	n
South Coast	261	38.7	-0.13	0.14	0.35	99
The Karoo	270	40.9	-0.03	0.15	0.82	99

Both the South Coast and Karoo clusters show no evidence of linear trend (*Table 48*). For the South Coast cluster, around 1939, the mean cessation day does pass near the bottom boundary as the GAM 95% confidence interval has shifted upward. Between 1964 and 1969, the confidence shifts completely above the mean cessation day (*Figure 49, top*).

For the Karoo cluster, the GAM 95% confidence interval always covers the mean cessation day. Around 1934, the confidence interval does shift down and mean cessation day passes near the top boundary (*Figure 49, bottom*).

4.8.7 Comparison of Different Approaches to Estimate the End of Season

The linear and non-linear trends are compared for each cluster for all of the methods used to determine the end of season. For the linear trend, the calculated p-values (*Tables 43 - 48*) and confidence intervals of the gradient of the linear model (*Table 49*) are compared and for the non-linear trends, the years where the mean cessation day does not fall within the GAM 95% confidence interval are compared for each cluster (*Tables 50 - 52*).

Abbreviations of Methods:

- 'Cum. Thr.' = 'Cumulative Threshold' (chapter 4.8.1)
- 'Mod. Def.' = 'Modified Definition' (chapter 4.8.2)
- 'Thr. Mean' = 'Threshold from Mean Daily Rainfall' (chapter 4.8.3)
- 'Thr. Pr(0)' = 'Threshold from Probability of Zero' (chapter 4.8.4)
- 'Grad. Mean' = 'Gradient from Mean Daily Rainfall' (chapter 4.8.5)
- 'Grad. Pr(0)' = 'Gradient from Probability of Zero' (chapter 4.8.6)

Table 49: Comparing the 95% confidence intervals for the gradient of the linear models for the end of season for the summer rainfall clusters (the South Coast and Karoo). Confidence intervals in bold represent confidence intervals showing evidence of change.

	Cum. Thr.		Mod. Def.		Thr. Mean		Thr. Pr(0)		Grad. Mean		Grad. Pr(0)	
	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL	LL	UL
South Coast	-0.10	0.70	-0.16	0.20	-0.48	0.25	-0.53	0.18	-0.13	0.45	-0.40	0.14
The Karoo	**-0.82**	**-0.05**	-0.26	0.15	-0.48	0.35	-0.23	0.48	-0.41	0.15	-0.32	0.25

There is only one method that shows evidence of linear change and that is for the Karoo cluster - when using a cumulative threshold (chapter 4.8.1). The other confidence intervals for the Karoo cluster are mostly on the negative side (*Table 49*). There is some evidence of linear trend to an earlier end of season for this cluster. The South Coast cluster shows no evidence of linear change with contradicting confidence intervals.

Table 50: South Coast - Years where the GAM 95% confidence interval does not Contain the mean cessation day. Years in italics represent the mean cessation day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean cessation day is completely below the confidence interval indicating an earlier end to the season relative to the mean. 'Later' refers to years where the mean cessation day is completely above the confidence interval indicating a later end to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +	
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier	Later
Cum. Thr.	1932 to 1937		1950 to 1955			1972 to 1977		
Mod. Def.	*1932*					1978 to 1981		
Thr. Mean								
Thr. Pr(0)								
Grad. Mean								
Grad. Pr(0)		*1939*		1964 to 1969				

As there are no multiple overlapping results from the methods no conclusion can be drawn for this cluster other than that there has been no non-linear changes to the cessation of the season other than natural variation (*Table 50*).

Table 51: The Karoo - Years where the GAM 95% confidence interval does not Contain the mean cessation day. Years in italics represent the mean cessation day within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Earlier' refers to years where the mean cessation day is completely below the confidence interval indicating an earlier end to the season relative to the mean. 'Later' refers to years where the mean cessation day is completely above the confidence interval indicating a later end to the season relative to the mean.

	1918 to 1945		1946 to 1970		1971 to 1995		1996 +	
	Earlier	Later	Earlier	Later	Earlier	Later	Earlier	Later
Cum. Thr.		*1921*					2010	
Mod. Def.		*1932*			*1991*			
Thr. Mean								
Thr. Pr(0)			*1957*					
Grad. Mean				*1969*				
Grad. Pr(0)	*1934*							

Similar to the South Coast cluster, no conclusion can be made for the Karoo cluster, other than there is no evidence for non-linear change. This is indicated by the GAM 95% confidence interval not leaving the mean cessation day across multiple of the methods (*Table 51*).

4.9 Length of the Rainfall Season

The length of the rainfall season is computed over the 100 years (1918 to 2017) in order to track any changes in length.

4.9.1 Winter Rainfall Clusters

Estimated length for the winter rainfall clusters (Mediterranean).

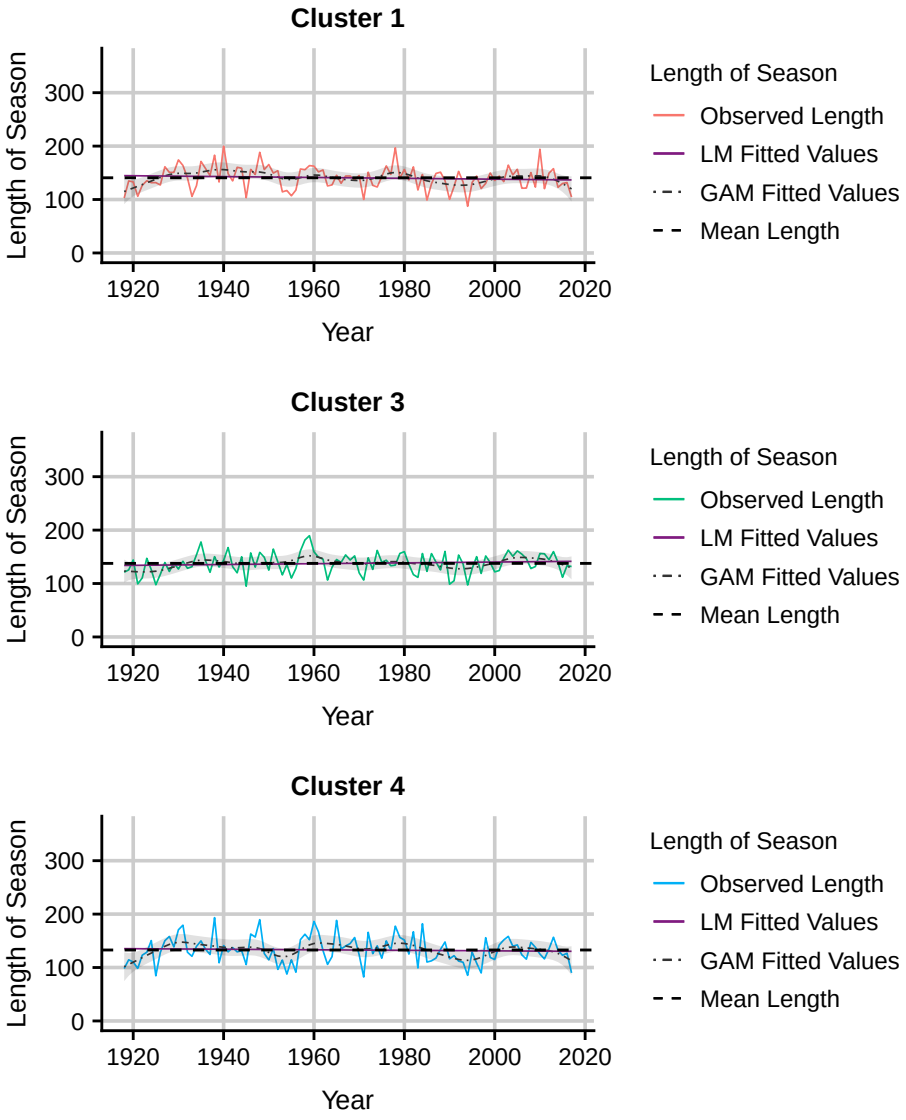


Figure 50: Estimated length of rainfall season (1918 to 2017).

Table 52: Summary of the linear model fitted to the estimated length dates for winter rainfall clusters (1, 3 and 4).

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value
South Mediterranean	141	21.9	-0.08	0.08	0.29
Central Mediterranean	138	19.3	0.08	0.07	0.25
West Mediterranean	133	25.2	-0.05	0.09	0.57

None of the clusters show any evidence of linear trend (*Table 52*). Looking at the non-linear trend, the GAM 95% confidence interval for South Mediterranean cluster is initially completely below the mean length between 1918 and 1921. The confidence interval quickly rises above the mean length between 1937 and 1941. The mean length passes near the top boundary around 1978. The confidence interval then shifts downward until the mean length passes completely below the confidence interval between 1991 and 1994. Lastly, the mean length passes through the top boundary of the confidence interval around 2017 (*Figure 50, top*).

The confidence interval for the Central Mediterranean cluster is also initially below the mean length between 1920 and 1926. The confidence interval then rises until it is completely above the mean length at 1959. Around 1993, the confidence interval has shifted downward and the mean length passes near the top boundary. Lastly, around 2006, the confidence interval has shifted upward and the mean onset day passes through the bottom boundary (*Figure 50, middle*).

The same as the above clusters, the confidence interval for the West Mediterranean cluster is also initially completely below the mean length between 1918 and 1921. The confidence interval then shifts upward and the mean length passes through the bottom boundary around 1931. The confidence interval then shifts downward until the mean length passes through the top boundary around 1953. Then around 1961, the mean length passes through the bottom boundary. Again around 1979, the mean length passes through the bottom boundary. Between 1991 and 1996, the confidence interval has shifted completely below the mean length (*Figure 50, bottom*).

Table 53: Years where the GAM 95% confidence interval does not contain the mean length for the winter rainfall clusters. Years in italics represent the mean length within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Shorter' refers to years where the mean length passes below the confidence interval indicating a shorter length relative to the mean. 'Longer' refers to years where the mean length passes above the confidence interval indicating a longer length relative to the mean.

	1918 to 1940		1941 to 1960		1961 to 1980		1981 to 2000		2001 to 2017	
	Shorter	Longer	Shorter	Longer	Shorter	Longer	Shorter	Longer	Shorter	Longer
South Mediterranean	1918 to 1921	1937 to 1941		*1978*			1991 to 1994		*2017*	
Central Mediterranean	1920 to 1926			1959			*1993*			*2006*
West Mediterranean	1918 to 1921	*1931*	*1953*			*1961*	1991 to 1996	*1979*		

Initially the confidence interval from all the clusters is completely below the mean length (*Table 53*). During the mid 1990s, all of the clusters show a trough. This indicates that between 1918 to roughly 1925, there was a shorter season length. Thereafter, there was some natural variation until the mid 1990s where again there appears to have been a shortening to the length for this period.

4.9.2 Summer Rainfall Clusters

Estimated length for the summer rainfall clusters (the South Coast and Karoo).

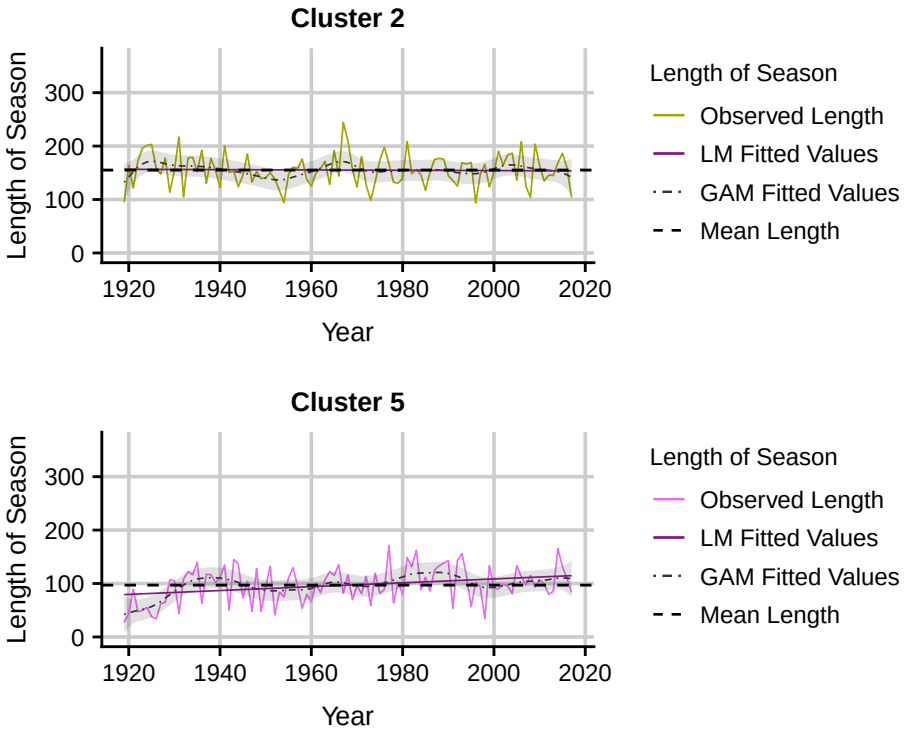


Figure 51: Estimated length of rainfall season (1918 to 2017).

Table 54: Summary of the linear model fitted to the estimated length dates for the summer rainfall clusters (the South Coast and Karoo).

Cluster	Mean Onset Day	Std. Deviation	Gradient	Std. Error	p-value
South Coast	155	30.2	-0.03	0.11	0.78
The Karoo	97	33.1	0.36	0.11	0.00

The South Coast cluster shows no evidence of linear trend (*Table 54*). Looking at the non-linear trend, around 1925, the GAM 95% confidence interval rises and the mean length passes through the bottom boundary. Around 1953, the confidence interval has shifted downward and the mean length passes through the top boundary. Around 1967, the confidence interval has shifted upward and the mean length passes through the bottom boundary. The average length appears stable thereafter (*Figure 51, top*).

The Karoo cluster shows strong evidence of positive linear trend with a p-value of 0.00 (*Table 54*). For the non-linear trend, initially the GAM 95% confidence interval is completely below the average length from 1919 to 1928. Between 1982 and 1990, the confidence interval is completely above the mean length. Thereafter, the mean length is stable (*Figure 51, bottom*).

Table 55: Years where the GAM 95% confidence interval does not contain the mean length for the summer rainfall clusters. Years in italics represent the mean length within the confidence interval, passing within 5 days of the boundary of the confidence interval. 'Shorter' refers to years where the mean length passes below the confidence interval indicating a shorter length relative to the mean. 'Longer' refers to years where the mean length passes above the confidence interval indicating a longer length relative to the mean.

	1918 to 1940		1941 to 1960		1961 to 1980		1981 to 2017	
	Shorter	Longer	Shorter	Longer	Shorter	Longer	Shorter	Longer
South Coast		*1925*	*1953*			*1967*		
The Karoo	1919 to 1928							1982 to 1990

The South Coast cluster only appears to have natural variation to its length in season whilst the Karoo cluster seems to have had an increasing length in season. This is highlighted by the linear model (*Table 54*) as well as the rising of the confidence interval (*Table 55*).

4.10 Comparison of the Clusters and the Methods to Estimate Start and End of Season

Six different methods have been used to estimate the start and end of season for all the clusters.

Abbreviations of Methods:

- 'Cum. Thr.' = 'Cumulative Threshold'
- 'Mod. Def.' = 'Modified Definition'
- 'Thr. Mean' = 'Threshold from Mean Daily Rainfall'
- 'Thr. Pr(0)' = 'Threshold from Probability of Zero'
- 'Grad. Mean' = 'Gradient from Mean Daily Rainfall'
- 'Grad. Pr(0)' = 'Gradient from Probability of Zero'

Start of season:

The average start of season for the clusters are 26 April, 13 October, 29 April, 29 April and 15 December for clusters 1, 2, 3, 4 and 5, respectively (*Table 56*). The average start of season day for the winter rainfall clusters (1, 3 and 4) are all very similar (within four days of each other) which is to be expected and further supports the final clusters chosen (*Table 12*).

Table 56: Mean start of season days. The Mediterranean Clusters represent days since 1 January and the South Coast and Karoo clusters represent days since 1 July.

	Clusters				
	1	2	3	4	5
Cum. Thr.	101	109	108	97	194
Mod. Def.	121	78	126	133	122
Thr. Mean	119	103	119	113	157
Thr. Pr(0)	111	69	110	120	145
Grad. Mean	111	134	118	113	173
Grad. Pr(0)	135	133	134	139	212
Mean	116.3	104.3	119.2	119.2	167.2

The summer rainfall clusters (the South Coast and Karoo) have a much higher standard deviation across the methods than the winter rainfall clusters (*Table 57*). The average standard deviations for the South Coast and Karoo clusters are 46.4 and 53.8 respectively, compared to the average standard deviations of the winter rainfall clusters which are 28.8, 26.8 and 30.9 for clusters 1, 3 and 4, respectively. This is expected as there is a much more defined rainfall season in the winter rainfall clusters.

Table 57: Standard deviation of start of season days.

	Clusters				
	1	2	3	4	5
Cum. Thr.	27.5	31.7	25.1	29.9	55.5
Mod. Def.	23.7	29.4	22.8	28.6	52.5
Thr. Mean	26.6	64.3	25.9	34.6	63.7
Thr. Pr(0)	29.4	44.9	25.1	28.7	61.9
Grad. Mean	27.5	46.8	26.4	28.4	57.1
Grad. Pr(0)	38.2	61.3	35.5	35.4	32.1
Mean	28.8	46.4	26.8	30.9	53.8

Clustering using a threshold on the annual cumulative rainfall produces lower estimated start of season dates on average for the winter rainfall clusters. The start of season estimates for most of the other methods are slightly lagged. This is because the start of season estimated by the other methods find a date already after the season has started. The estimates obtained when using the GAM models on average have a higher variance than the other two methods. Even though the hope was to reduce some variation by fitting a smooth model to the data, there is a lot of variability from year to year resulting in different start of season estimates.

End of Season:

The average end of season for the clusters are 14 September, 18 March (year + 1), 14 September, 8 September and 22 March (year + 1) (*Table 58*), respectively. Again, the end of season for the winter rainfall clusters are all very similar supporting the final clusters chosen.

Table 58: Mean end of season days. The Mediterranean clusters represent days since 1 January and the South Coast and Karoo clusters represent days since 1 July.

	Clusters				
	1	2	3	4	5
Cum. Thr.	247	216	236	234	245
Mod. Def.	245	265	243	244	244
Thr. Mean	245	271	249	242	270
Thr. Pr(0)	281	276	287	271	291
Grad. Mean	260	271	261	254	271
Grad. Pr(0)	263	261	263	262	270
Mean	256.8	260.0	256.5	251.2	265.2

The summer rainfall clusters have a higher standard deviation across the methods (*Table 59*). The average standard deviations for the South Coast and Karoo clusters are 43.7 and 42.5 respectively. The average standard deviations for clusters 1,3 and 4 are 35.6, 34.6 and 37.4 respectively. However, this is not as big a difference as the start of season standard deviations.

Table 59: Standard deviation of end of season days.

	Clusters				
	1	2	3	4	5
Cum. Thr.	48.6	56.9	40.4	48	51.2
Mod. Def.	23.9	25.6	23.6	30.2	29.8
Thr. Mean	29.4	51.2	32.2	37.5	50.4
Thr. Pr(0)	38.5	48.3	35.0	36.2	42.6
Grad. Mean	35.7	41.7	35.8	37.9	40.3
Grad. Pr(0)	37.4	38.7	40.6	34.6	40.9
Mean	35.6	43.7	34.6	37.4	42.5

4.11 Summary

The final clusters determined match well the underlying climate structure of the Western Cape. It is very difficult to conclude changes in start and end of season day. This is due to the huge natural variation of daily rainfall. However, some notable changes have occurred.

5. Conclusions

5.1 An Overview

30 weathers stations within the Western Cape are clustered according to annual seasonal patterns. Various methods are then used to determine start and end of season days for each year. These start and end of season days are then modeled over time using linear and generalized additive models to see whether there has been any changes in seasonality between 1918 and 2017. Lastly, the length of the season is analyzed based on the start/end of season days.

5.2 Summarising the Results

The final chosen clusters match the underlying climate structure of the Western Cape perfectly. Even though five clusters are chosen and there are only three different regions within the Western Cape, three of the clusters fit into the Mediterranean region. These three clusters differ in the amount of rainfall they receive, although their seasonal profiles are very similar. The clusters highlighted the strong winter rainfall season of the Mediterranean region and the fairly aseasonal rainfall of the Karoo and South Coast.

For most of the clusters, no evidence of changes in start date, end date or length are found. There are periods with earlier and later starts, but no consistent shift over time.

Looking at the start of season for the winter rainfall regions, there is evidence of linear changes for the Central and West Mediterranean clusters with an earlier start to the season. Based on the non-linear trend, there did appear to be a shift to an earlier start around 1945, with the start shifting back thereafter. For the Central Mediterranean cluster, there is evidence that there was a shift to an earlier start around 1952 that then re-stabilized. Around the early 1970s, there is evidence that the start of season was shifting later to this point and then re-stabilized afterward.

Looking at the end of season for the winter rainfall regions, there is evidence of linear changes for the South Mediterranean and West Mediterranean clusters - showing an earlier end to the season. Based on the non-linear trend, for the South Mediterranean cluster, there is strong evidence that there was a shift to a later end of season up until 1930, thereafter, the start of season returned back to the mean. All the other clusters resulted in no conclusive evidence for non-linear changes.

For the start of season for the summer rainfall regions (South Coast and the Karoo), there is strong evidence of linear trend for the Karoo cluster with the start of season changing to earlier in the year. This result is emphasized by the non-linear trend as evidence for shifts to an earlier start of the season are found around the early 1980s and around 2013. Based on the non-linear trend for the South Coast cluster, evidence is shown that there has been a shift to a later start around 1934, thereafter, the start returned back to the mean.

For the end of season for the summer rainfall regions, there is evidence of a linear change to an earlier end of season for the Karoo cluster. No non-linear changes are found.

The winter rainfall clusters seem to have a similar history of length of season. All of the clusters experienced shorter season length in the early 1920s and 1990s. The South Coast cluster appears to have had no change in length of season whilst the Karoo cluster shows strong evidence that there has been an increasing length of season over time.

It is important to note that the South Coast and Central Mediterranean clusters have much smaller sample sizes than the other three clusters and so, their results should be interpreted with caution.

Even though it has proven difficult to pick up on any strong shifts in the rainfall season, in the Western Cape - extreme events are likely to be on the rise. As such, it is important that this area continues to do research on this subject and to do its best to be prepared for future extreme events (particularly drought), by saving more water now.

5.3 Future Work and Suggestions

Many different methods can be used to estimate the start and end of season. Methods using the rainfall amount have been used in this thesis. Other papers have made use of the 'number of wet days' as opposed to the rainfall amount. This removes some of the variability of the rainfall amount.

5.4 General Thoughts

Daily rainfall data is a very difficult data set to work with due to the high percentage of zeros and the highly variable nature of rainfall. Many other studies have tried to estimate the start and end of season for various locations and have produced similar results in terms of high variation of estimates. The 'start' and 'end of season' will shift depending on how you define them. However, as long as the method of start/end of season estimation is used consistently over time, the general patterns of start/end of season days can be found.

References

Abrahams, W. 2019. 'Is the rainy season shifting?', *CSAG blog*, 8 April.Available: http://www.csag.uct.ac.za/2019/04/08/is-the-rainy-season-shifting/.

Alam, M., Paul, S. 2019. 'A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh', *Journal of Applied Statistics*, pp. 1-22.

Azur, M., Stuart, E., Frangakis, C., Leaf, P. 2012. 'Multiple Imputation by Chained Equations: What it is and how does it work?', *International Journal of Methods in Psychiatric Research*, vol. 20(1), pp. 40-49.

Barnett, T., Adam, J., Lettenmaier, D. 2005. 'Potential impacts of a warming climate on water availability in snow-dominated regions', *Nature Publishing Group*, vol. 438, pp. 303-309.

Conway, D., Dalin, C., Landman, W.A. and Osborn, T.J. (2017). 'Hydropower plans in eastern and southern Africa increase risk of concurrent climate-related electricity supply disruption'. *Nature Energy*, vol. 2, pp. 946-953.

Cowling, R., Esler, K., Rundel, P. 1999. 'Namaqualand, South Africa - an overview of a unique winter-rainfall desert ecosystem', *Plant Ecology*, vol. 142(1), pp. 3-21.

Dourte, D., Fraisse, C., Bartels, W. 2015. 'Exploring changes in rainfall intensity and seasonal variability in the Southeastern U.S.: Stakeholder engagement, observations, and adaptation', *Climate Risk Management*, vol. 7, pp. 11-19.

du Plessis, J., Schloms, B. 2017. 'An investigation into the evidence of seasonal rainfall pattern shifts in the Western Cape, South Africa', *Journal of the South African Institution of Civil Engineering*, vol. 59(4), pp. 47-55.

Erni, B. 2018. *Splines and GAMs*, lecture notes, Advanced Regression STA5003W, Department of Statistical Sciences, University of Cape Town.

Grosjean, P., Ibanez, F. 2018. pastecs: Package for Analysis of Space-Time Ecological Series. R package version 1.3.21. https://CRAN.R-project.org/package=pastecs.

Hastie, T., Tibshirani, R. 1990. 'Generalized Additive Models', *Springer*.

Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, W., Medina-Elizade, M. 2006. 'Global temperature change', *Proceedings of the National Academy of Sciences*, vol. 103, pp. 14288-14293.

Heitjant, D., Little, R. 1991. 'Multiple Imputation for the Fatal Accident Reporting System', *Journal of the Royal Statistical Society*, vol. 40(1), pp. 13-29.

Johnson, R., Wicharn, D. 2007. 'Applied Multivariate Statistical Analysis, 6th ed.', *Springer*.

Jolliffe, I., Sarria-Dodd, D. 1994. 'Early detection of the start of the wet season in semiarid tropical climates of western Africa', *International Journal of Climatology*, vol. 21(10), pp. 1251-1262.

Kruger, A. 2006. 'Observed Trends in Daily Precipitation Indices in South Africa: 1910-2004', *Journal of Climatology*, vol. 26, pp. 2275-2285.

Kruger, A., Nxumalo, M. 2017. 'Historical rainfall trends in South Africa: 1921-2015' *Water South Africa*, vol. 43(2), pp. 285.

Laux, P., Kunstmann, H., Bardossy, A. 2008. 'Predicting the Regional Onset of the Rainy Season in West Africa', *International Journal of Climatology*, vol. 28(3), pp. 329-342.

Lobell, D., Schlenker, W., Costa_Roberts, J. 2011. 'Climate Trends and Global Crop Production Since 1980', *Science*, vol. 333(6042), pp. 616-620.

Lumsden, T., Schulze, R., Hewitson, B. 2007. 'Evaluation of potential changes in hydrologically relevant statistics of rainfall in Southern Africa under conditions of climate change', *Water South Africa*, vol. 35(5), pp. 649-656.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2018). 'cluster: Cluster Analysis Basics and Extensions'. R package version 2.0.7-1.

Mimmack, G. M., Mason, S. J. & Galpin, J. S. (2001). 'Choice of distance matrices in cluster analysis: defining regions'. *Journal of Climate*, vol. 14, pp. 2790-2797.

Murphy, B., Timbal, B. 2006. 'A review of recent climate variability and climate change in southeastern Australia', *International Journal of Climatology*, vol. 28(7), pp. 859-879.

New, M., Hewitson, B., Stephenson, D. 2006. 'Evidence of trends in daily climate extremes over southern and west Africa', *Journal of Geophysical Research: Atmospheres*, vol. 111(14).

Odenkule, T. 2004. 'Determining Rainfall Onset and Retreat Dates in Nigeria', *Journal of Human Ecology*, vol. 16, pp. 239-247.

Pohl, B., Macron, C., Monerie, P. 2017. 'Fewer rainy days and more extreme rainfall by the end of the century in Southern Africa', *Scientific Reports*, vol. 7, pp. 46466.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reason, C., Rouault, M. 2005. 'Links between the Antarctic Oscillation and winter rainfall over western South Africa', *Geophysical Research Letters*, vol. 32(7).

Rigby, R., Stasinopoulos, D. 2005. 'Generalized additive models for location, scale and shape,(with discussion)', *Applied Statististics*, vol. 54(3), pp. 507-554.

Rigby, R. Stasinopoulos, M. 2007. 'Generalized Additive Models for Location Scale and Shape (GAMLSS) in R', *Journal of Statistical Software*, vol. 23(7).

Schafer, J., Graham, J. 2002. 'Missing data: Our view of the state of the art', *Psychological Methods*, vol. 7, pp. 147-177.

Schenker, N., Taylor, J. 1996. 'Partially parametric techniques for multiple imputation', *Computational Statistics and Data Analysis*, vol. 22(4), pp. 425-446.

Stern, R. 1981. 'The start of the rains in West Africa', *International Journal of Climatology*, vol. 1, pp. 59-68.

South African Weather Service, Daily Rainfall Data, 30 Weather Stations, 1918 to 2017.

Trenberth, K. 1997. 'The Definition of El Niño', *Bulletin of the American Meteorological Society*, vol. 78(12), pp. 2771-2777.

van Buuren, S., Groothuis-Oudshoorn, K. 2011. 'mice: Multivariate Imputation by Chained Equations in R'. *Journal of Statistical Software*, vol. 45(3), pp. 1-67. URL https://www.jstatsoft.org/v45/i03/.

van Heerden, J., Hastenrath, S., Greischar, L. 1994. 'Prediction of the Summer Rainfall over South Africa', *Journal of Climate*, vol. 8(6), pp. 1511-1518.

Van Niekerk, A., Joubert, S. 2011. 'Input variable selection for interpreting high-resolution climate surfaces for the Western Cape', *Water South Africa*, vol. 37(3), pp. 271-279.

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer*, 2016.

Wood, S. 2017. 'Generalized Additive Models : An Introduction with R, Second Edition', *Chapman and Hall/CRC*.

Yaman, N. 2014. 'Modelling Precipitation of certain regions for Turkey via Hidden Markov Models', MSc thesis, Middle East Technical University, Ankara.

Yulei, H., Trivellore, R. 2008. 'On the Performance of Sequential Regression Multiple Imputation Methods with Non Normal Error Distributions', *Communications in Statistics - Simulation and Computation*, vol. 38(4), pp. 856-883.

Zhang, G., Nearing, M., Liu, B. 2005. 'Potential Effects of Climate Change on Rainfall Erosivity in the Yellow River Basin of China', *Transactions of the ASAE*, vol. 48(2), pp. 511-517.