Journal of Hydroinformatics

Journal of Hydroinformatics Vol 26 No 1, 162 doi: 10.2166/hydro.2023.261

Identifying the pathways of extreme rainfall in South Africa using storm trajectory analysis and unsupervised machine learning techniques

Rhys Phillips[a], Katelyn Ann Johnson[b,c], Andrew Paul Barnes[d] and Thomas Rodding Kjeldsen [a,b,*]

[a] Department of Architecture and Civil Engineering, University of Bath, Bath BA2 7AY, UK
[b] School of Engineering, University of KwaZulu-Natal, Durban, South Africa
[c] Centre for Water Resources Research, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[d] Department of Computer Science, University of Bath, Bath, UK
*Corresponding author. E-mail: trk23@bath.ac.uk

TRK, 0000-0001-9423-5203

ABSTRACT

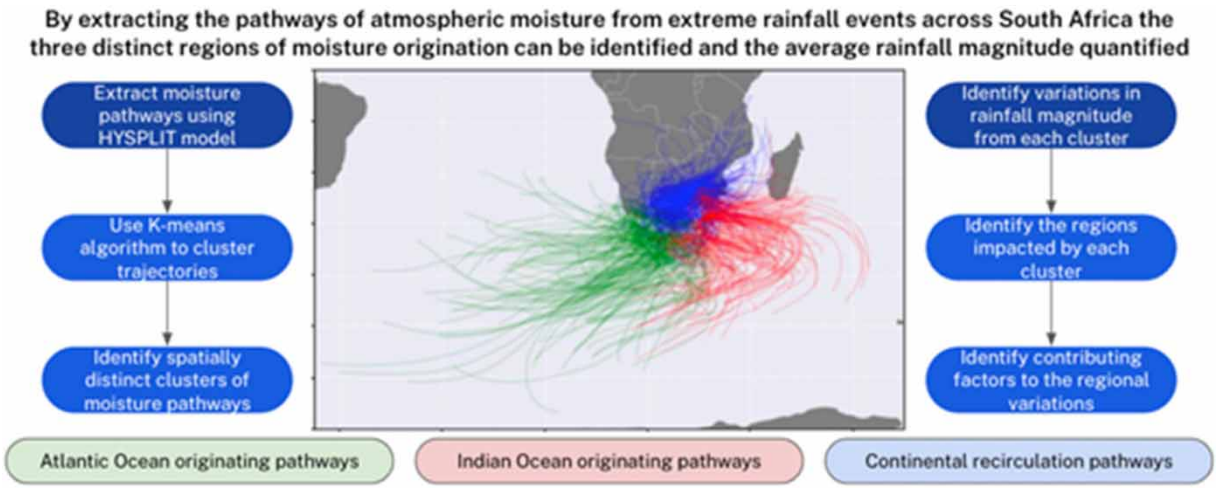
This study has utilised National Oceanic and Atmospheric Administration (NOAA) NCEP/NCAR Reanalysis 1 project meteorological data and the HYSPLIT model to extract the air parcel trajectories for selected historical extreme rainfall events in South Africa. The k-means unsupervised machine learning algorithm has been used to cluster the resulting trajectories, and from this, the spatial origin of moisture for each of the rainfall events has been determined. It has been demonstrated that rainfall events on the east coast with moisture originating from the Indian Ocean have distinctly larger average maximum daily rainfall magnitudes (279 mm) compared to those that occur on the west coast with Atlantic Ocean influences (149 mm) and those events occurring in the central plateau (150 mm) where moisture has been continentally recirculated. Further, this study has suggested new metrics by which the HYSPLIT trajectories may be assessed and demonstrated the applicability of trajectory clustering in a region not previously studied. This insight may in future facilitate improved early warning systems based on monitoring of atmospheric systems, and an understanding of rainfall magnitudes and origins can be used to improve the prediction of design floods for infrastructure design.

Key words: extreme rainfall, *k*-means clustering, South Africa, trajectories, unsupervised learning

HIGHLIGHTS

- Clustering extreme rainfall events based on origins and storm track.
- Observed differences in magnitude between event clusters.
- Depth of rainfall events (mm/day) from weather systems originating in the Indian Ocean is larger than other events.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Extreme rainfall events, and the resulting flooding, are a cause of particular concern to countries across Africa where increasing flood-induced economic impacts are predicted to be driven by climate change (di Baldassarre *et al.* 2010; Winsemius *et al.* 2016). For example, Aon (2020) reported that 39 of the 45 *Global Disasters* recorded in Africa in 2020 were attributed to flooding, and in the past decade, flooding has surpassed drought as the natural disaster affecting the greatest number of people on the African continent (Lumbroso 2020). Aside from the direct damage to human life and infrastructure, the impact of flooding on public health can be considerable. As well as, the disruption of medical services and the impacts of flooding can cause a drastic increase in vector-borne diseases (Ahern *et al.* 2005) and natural disasters resulting from extreme rainfall, which can exacerbate existing vulnerabilities in underserved and informally established, flood-prone, peri-urban areas communities (Khandlhela & May 2006).

As with many countries across the continent, deadly floods due to heavy rainfall have occurred across South Africa in recent years (le Maitre *et al.* 2019). In South Africa, the rate of urbanization has steadily increased (World Bank 2021), resulting in amplified pressures on existing services and infrastructure, negatively affecting infrastructure resilience while potentially exacerbating the impacts of any flood event; for example, as a result of the 77 flood events listed between 1980 and 2010, over 1,000 people were thought to have died (Zuma *et al.* 2012). Recent floods caused by heavy rainfall in 2011 alone killed more than 40 people and caused $51 million in damages nationwide (Mabuse 2021), while the 2019 Durban floods are thought to have killed 70 people and caused $45 million in damages (UNOOSA 2019).

South Africa is a semi-arid country, which experiences an uneven spatial distribution of rainfall, and there is a notable range in the total annual rainfall amount and the seasonal distribution of rainfall (Roffe *et al.* 2019), from approximately 250 mm in the west, at approximately 20° longitude, to over 1,000 mm in the east, at approximately 30° longitude (International Food Policy Research Institute (IFPRI) 2014). However, recent research carried out by the South African Weather Service (SAWS 2019) found that precipitation from the most intense rainfall events is increasing across the nation, while there was a decreasing trend in annual rainfall across most regions and an increase in annual rainfall in the southern interior. This highly variable rainfall is due in part to South Africa's location, being influenced by both the South Atlantic and the cold Benguela current on the west coast (Hahn *et al.* 2017), as well as the South Indian Ocean and the warm Agulhas current on the east coast (Jury 2015). In addition, the latitude of the southern tip of South Africa exposes the South Western Cape area to mid-latitude cyclones causing winter rainfall events to dominate in this region, in comparison to the majority of South Africa, which predominantly receives summer rainfall (Odoulami *et al.* 2020).

Published research has identified that the frequency and magnitude of extreme rainfall events have increased in some parts of South Africa (Ziervogel *et al.* 2014), and future rainfall conditions are projected to further intensify and become more extreme for many regions in the country (du Plessis & Burger 2015; de Waal *et al.* 2017). Increased frequency and magnitude of extreme events amplify potential flood risks. Therefore, understanding the climate drivers and origins of extreme rainfalls is becoming increasingly important to researchers and practitioners in updating design strategies for hydraulic infrastructure in South Africa (Schulze & Schütte 2019). This is critical for water resources planning and design of future infrastructure, maintaining and upgrading existing hydraulic structures, and mitigating risks to current urban stormwater drainage systems in a rapidly developing country.

Recognising the importance of developing an understanding of extreme rainfall events in South Africa, this study focuses on investigating the distinct regions of moisture origin for extreme rainfall in South Africa. This has been achieved by utilising the HYSPLIT (Hybrid Single Particle Lagrangian Integrated Trajectory) model (Stein *et al.* 2015) to extract the storm trajectories, defined here as the pathways that moisture followed through the atmosphere, for a series of historical maximum magnitude rainfall events recorded at weather monitoring stations across South Africa in the years 1950–2010. These trajectories have then been clustered using unsupervised machine learning techniques to determine the regions of moisture origin and the differences in event rainfall magnitude that they cause, in different regions of South Africa. The HYSPLIT model has been used extensively in other studies as a tool for tracing atmospheric moisture, for example, in studies of atmospheric chemistry origins (Xia *et al.* 2020; Ma *et al.* 2021), simulation of dust storms (McGowan & Clark 2008; Ashrafi *et al.* 2014), and volcanic ash (Hurst & Davis 2017). Notably, several researchers have used the HYSPLIT system when defining weather patterns and their synoptic conditions, focusing primarily on the classification of extreme events (Santos *et al.* 2018; Barnes *et al.* 2020; Karozis *et al.* 2021).

Described in this article are the data used and the techniques used to validate this data, followed by the methodology applied and the results of the clustering. These results are discussed, and conclusions are presented.

2. DATA

The database of events contains the single maximum rainfall magnitude records selected from the annual maximum series at 378 rainfall recording stations located throughout South Africa as shown in Figure 1. The spatial distribution of rainfall monitoring stations across the country (Figure 1) shows greater density of monitoring stations in the east of the country where the data reveal generally higher maximum rainfall events. Both the size and the colour of the location markers indicate maximum rainfall intensity with larger dots corresponding to larger rainfall magnitudes.

Each event is characterised by the longitude and latitude positions of the monitoring station, the maximum recorded rainfall (mm), the date occurrence, and the first and last year of continuous records that exist and have been considered for each station. The data was made available from a study updating the probable maximum precipitation (PMP – the theoretical upper limit for rainfall used for engineering design purposes) values for South Africa (Johnson & Smithers 2020), which considered 1,629 rainfall monitoring stations with at least 40 years of record available and selected a spatially representative sample of extreme rainfall events that consisted of continuously recorded data. The PMP events are typically required for engineering design of critical infrastructure where failure is catastrophic such as large reservoirs and nuclear installations (Wang *et al.* 2019. The database of events used for this study is therefore assessed to be suitable for the study of extreme rainfall events.

In addition to the date, longitude and latitude coordinates requires as an input to HYSPLIT, and the model also requires altitude, time of day, and length of extraction values for each trajectory calculation. For each of the 376 events, the trajectories were extracted for four times evenly spaced throughout the day (00:00, 08:00, 16:00, and 24:00), and these trajectory calculations were initiated at six altitudes (10, 410, 810, 1,210, 1,610, and 2,010 m), which corresponds to the altitude range in which moisture is expected to be found in the atmosphere (Wallace & Hobbs 2006). Thus, each rainfall day is covered by 24 trajectories. A length of 5 days was initially selected for the duration of trajectory back-calculation to ensure that sufficient

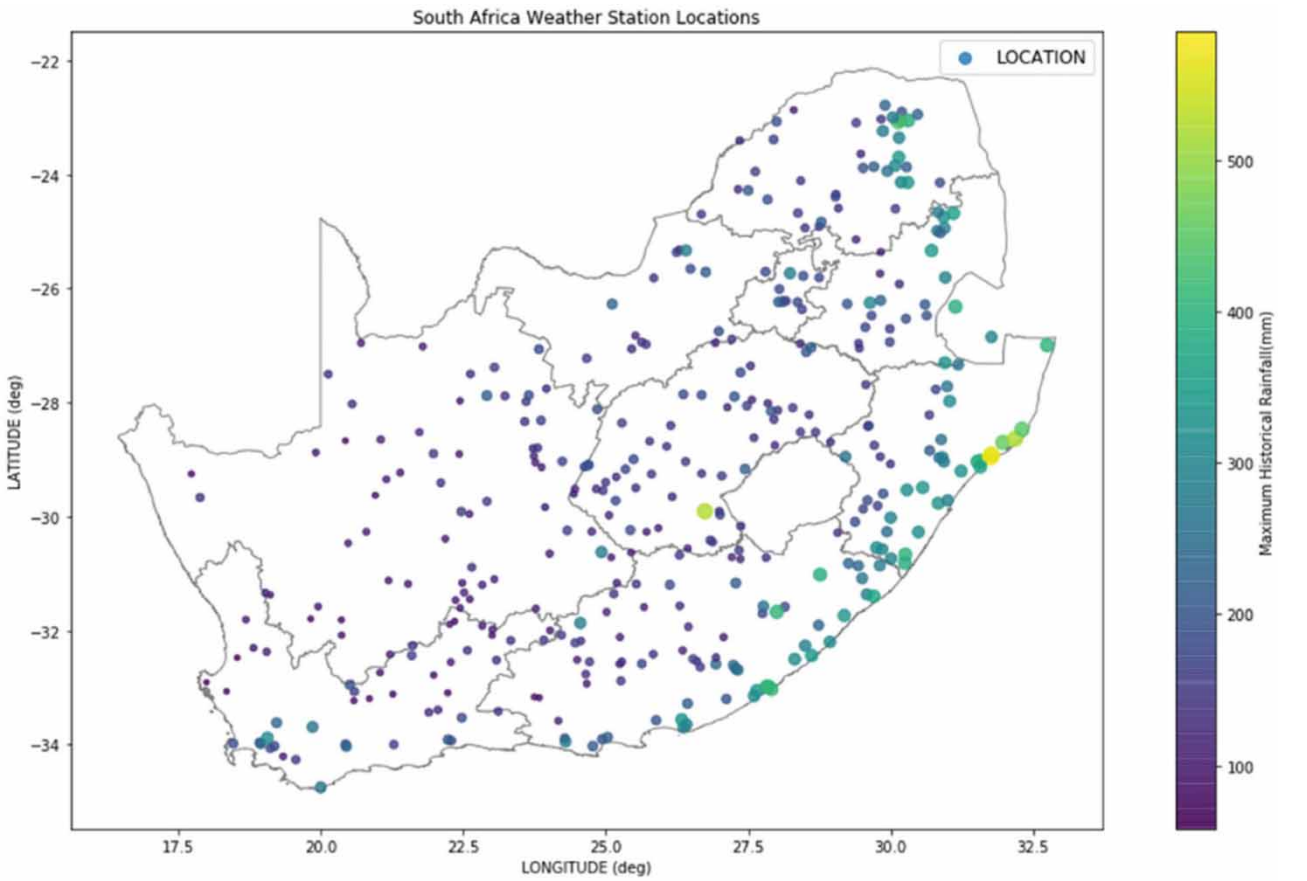


Figure 1 | Map of South Africa showing the location of the rainfall gauging stations considered in this study and the associated maximum magnitude of rainfall for the event recorded.

data would be gathered for the *k*-means clustering process while recognising that the output of an HYSPLIT analysis can be simplified to a $(n \times 2)$ matrix where each row contains the latitude and longitude positions of the air parcel at that interval and the number of rows n is the number of hours for which the back-calculation has been initialised – 5 day trajectories can therefore be shortened during analysis.

The 4 times and 6 altitudes selected yielded 24 trajectories representing each of the 376 individual events resulting in a total of $24 \times 376 = 9{,}024$ trajectories. Due to a limitation of the HYSPLIT model, events that occurred on the last day of the month were not able to generate the trajectory specified for 24:00 at each of the six altitudes (HYSPLIT is unable to 'roll over' to the next day when the last day of a month is selected as the day of initiation as it requires accessing two different meteorological data files during the same calculation). This has resulted in six trajectories for each of the 15 events that occurred on the final day of the month not being generated, and therefore, a total of 8,934 trajectories being generated. As this represents less than 1% of the total trajectories being lost and the affected events still have trajectories generated for all altitudes, they have been included in the clustering process, and it is not likely that this will cause any significant impact to the results.

3. METHODOLOGY

This section outlines the methodology used to obtain the storm trajectories and perform the clustering analysis which form the basis of the subsequent meteorological interpretation.

3.1. Event storm trajectories

Samples of trajectories with 1-, 2-, 3-, and 4-day lengths were plotted to visually inspect the pathways of a random sample of trajectories and ensure that they followed pathways that are considered credible from a meteorological perspective. This was also done to gauge the length of trajectories that should be considered when defining distinct event types through clustering of the trajectories. Too short a trajectory considered may lead to clusters that are not spatially different, as there will be necessarily very little spatial difference between the location of events (as they are evenly spread throughout South Africa), and too long a trajectory considered would be nonsensical for this study of moisture origin as it would entail considering trajectories that have been influenced by multiple atmospheric systems and would likely not be a good representation of the relevant moisture pathways. In addition, the larger errors inherent in longer trajectories would be potentially introduced. It is important to note that HYSPLIT exports air parcel trajectories, and it must be inferred the distance the moisture in this air parcel has travelled prior to being deposited as rainfall during the rainfall event in question. For example, if a trajectory was generated from an event in South Africa and showed that 5 days previously, the air parcel had passed over South America, this does not necessarily mean that the moisture deposited as rainfall during this event was carried over South America. In this case, it is judged to be far more likely that the moisture was taken into the atmosphere over the South Atlantic.

From the initial test plots shown in Figure 2, it can be seen that 1- and 2-day trajectories appear to be predominantly located close to the landmass of South Africa, while after 3 days, the same trajectories are far more spread out. After 4 days, some trajectories have travelled very large distances, predominantly over the Atlantic Ocean (1 in 10 trajectories shown for clarity).

Figure 2(a) and 2(b) show that 1-day and 2-day trajectories are centred around the African continent and will likely not give the spatial difference required to perform an effective cluster analysis and will therefore be unlikely to give an indication of the spatial origin of moisture for the extreme rainfall events considered. Figure 2(d) also demonstrates that by considering trajectories of lengths 4 days and above, the clustering process will be considering elements of trajectories that are being influenced by other atmospheric systems than those generating the extreme rainfall event under consideration.

Three-day trajectory lengths (Figure 2(c)) were therefore chosen for the analysis. The longitude and latitude coordinates of each trajectory, for each hourly interval within the 72 h previous to the rainfall event, were extracted and stored in individual vectors for ease of plotting. These vectors were then concatenated into a single matrix containing all position data for all trajectories to be used as input for the *k*-means clustering.

3.2. Trajectory clustering

Clustering the trajectories extracted into visually distinct groups necessitated the selection of an unsupervised machine learning technique, as this was an exploratory analysis that was efficient and appropriate for the task. Many unsupervised clustering techniques exist, but this study adopted the *k*-means algorithm (Hartigan & Wong 1979) due to its inherent simplicity and efficiency when considering clusters (Cui *et al.* 2021), being a Euclidean distance minimisation algorithm. The *k*-means algorithm was also adopted for use in other storm trajectory classification studies, notably by both Santos *et al.*

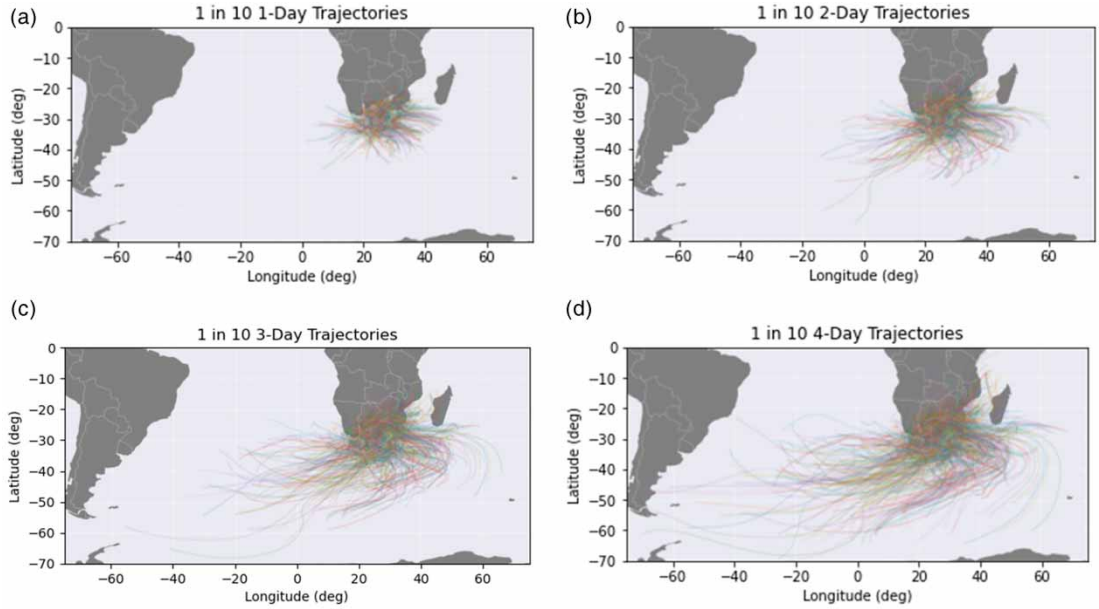


Figure 2 | One in 10 trajectories for (a, upper left) 1-day trajectories; (b, upper right) 2-day trajectories; (c, bottom left) 3-day trajectories; and (d, bottom right) 4-day trajectories.

(2018) and Barnes *et al.* (2019). The *k*-means algorithm also benefits from being a simple algorithm to understand which reduces the 'black-box' effect and associated uncertainty (Evans *et al.* 2019) when using the algorithm, furthering confidence in the results. The basic procedure that *k*-means carries out has been summarised below and adapted from Hartigan & Wong (1979):

1. Obtain a matrix of M points in N dimensions.
2. Select K initial cluster centres.
3. Assign all M points to the cluster's centre closest to them in Euclidian space.
4. Redefine cluster centres to be the average of the points contained within the cluster.
5. Re-allocate points to the nearest adjacent cluster centre.
6. Repeat steps 4 and 5 until all points remain in the same cluster

In the context of HYSPLIT trajectories, each of the M points is a single trajectory made up of a series of n pairs of longitude, and latitude coordinates representing the position of the air parcel at each 1-h interval. As each trajectory contains $n = 73$ pairs of coordinates which were flattened to form a vector, $N = 73 \times 2 = 146$ and the resulting matrix used for clustering is of dimensions $(N \times M)$. When considering 2D data, the cluster centre can be easily visualised as the mean average point of all the points within the cluster, whereas when considering trajectories, the cluster centre becomes the 'average' trajectory of a cluster.

The trajectories generated by the HYSPLIT model contain a number of parameters that can be analysed using the *k*-means algorithm. From the HYSPLIT output, the longitude, latitude, and altitude are listed at hourly intervals for each trajectory. In this study, only longitude and latitude have been considered for clustering, despite the availability of altitude data and the capacity of *k*-means to cluster data with different units. This is primarily because the altitude of trajectories is to an extent arbitrary, being chosen during this study. This is necessary as it is not possible to determine the altitude at which the actual precipitation is formed using HYSPLIT. Further to this, the previous work (Barnes *et al.* 2020) has indicated that altitude is a poor variable to consider when clustering as it adds dimensionality to the data and thus reduces the efficiency of the clustering algorithm when compared to the simpler case of using only longitude and latitude. Furthermore, the clustering of longitude and latitude is sufficient to produce visually different, spatially coherent trajectory clusters as demonstrated in this project and previous work (Tan *et al.* 2018).

When determining the optimum number of clusters of trajectories, the silhouette score was calculated, which considers both within cluster and out of cluster error (Rousseeuw 1987). The silhouette score measures the proximity of a trajectory to its allocated cluster centre (to be minimised) and proximity to the other cluster centres (to be maximised) and has possible values in the range [−1,1] with scores closer to 1 which indicate more distinct clusters being formed and scores close to 0 which indicate overlapping clusters (SciKit Learn 2020).

The silhouette score can be calculated as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}} \tag{1}$$

where i represents a data point (in this case a trajectory), $a(i)$ is the average distance from the data point to the other points within the cluster, and $b(i)$ represents the average distance between the data point and the data points in the adjacent cluster. For a more detailed overview of the silhouette score's derivation, see the study by Rousseeuw (1987). This is considered a more rigorous method of determining the optimum number of clusters as the identified weather generating systems are spatially distinct (originating from the Atlantic Ocean, Indian Ocean, and continentally), and so, consideration of the difference between clusters should be given. Figure 3 shows the silhouette score plotted against the number of clusters. While even low numbers of clusters produce relatively low silhouette scores, this is to be expected due to the complex and intertwined nature of the trajectories displayed in Figure 2, which are likely to result in overlapping clusters.

From Figure 3, an optimal number of clusters of three can be determined. Whilst both $K = 1$ and $K = 2$ yield higher silhouette scores; these are to be disregarded. Clustering into a single cluster will by definition yield a maximum silhouette score and the fact that the rate of change of the silhouette changes little between $K = 1$ and $K = 3$ demonstrates that while $K = 2$ would be a mathematically efficient solution, some insight may be lost. Choosing $K = 3$, after which there is a dramatic change in the rate of change of the silhouette score, will yield only marginally less mathematically optimal results but give a greater number of clusters which will allow for more insight into the underlying weather event generating systems.

In addition, as discussed earlier, rainfall in South Africa is dominated by systems originating in three different regions: the Atlantic Ocean, the Indian Ocean, and continental recirculation. Therefore, this study has opted to use three clusters for the primary investigation of extreme rainfall events as this is supported by both the data analysis and the meteorological considerations.

Figure 3 | Silhouette score plotted against number of clusters, K.

4. RESULTS

The results of the trajectory clustering process are detailed in this section. First, the three distinct clusters are detailed and the events from which the trajectories were initially generated are allocated to a trajectory. The rainfall magnitude distributions of each cluster are then analysed, as well as the spatial distribution of events allocated to each cluster is investigated.

4.1. Clustering process

Figure 4(a)–4(c) show the trajectories from the three clusters (clusters 0, 1, and 2) plotted individually with their respective cluster centres (plotted in gold), while Figure 4(d) shows the three clusters combined without their cluster medians. Where trajectories have been plotted, a random sample of one in every 10 trajectories has been shown for visual clarity.

From Figure 4(a)–4(c), it can observed that the clustering process has resulted in visually distinct groupings of trajectories that appear to originate from three distinct geographical regions:

- Cluster 0: Indian Ocean
- Cluster 1: Atlantic Ocean
- Cluster 2: Continental Recirculation

4.2. Event allocation

For each rainfall event, 24 trajectories were generated at various altitudes throughout the atmosphere and times of day. A simple method of allocating each event to the cluster in which the most trajectories from that event have been allocated has been adopted, and the efficiency of this method has been investigated. Figure 5 shows that when clustering the trajectory data using $K = 3$, the majority of events have all associated trajectories allocated to the same cluster. Furthermore, Figure 5 shows that the majority of events can be allocated to a cluster which contains more than 50% of the trajectories of that event.

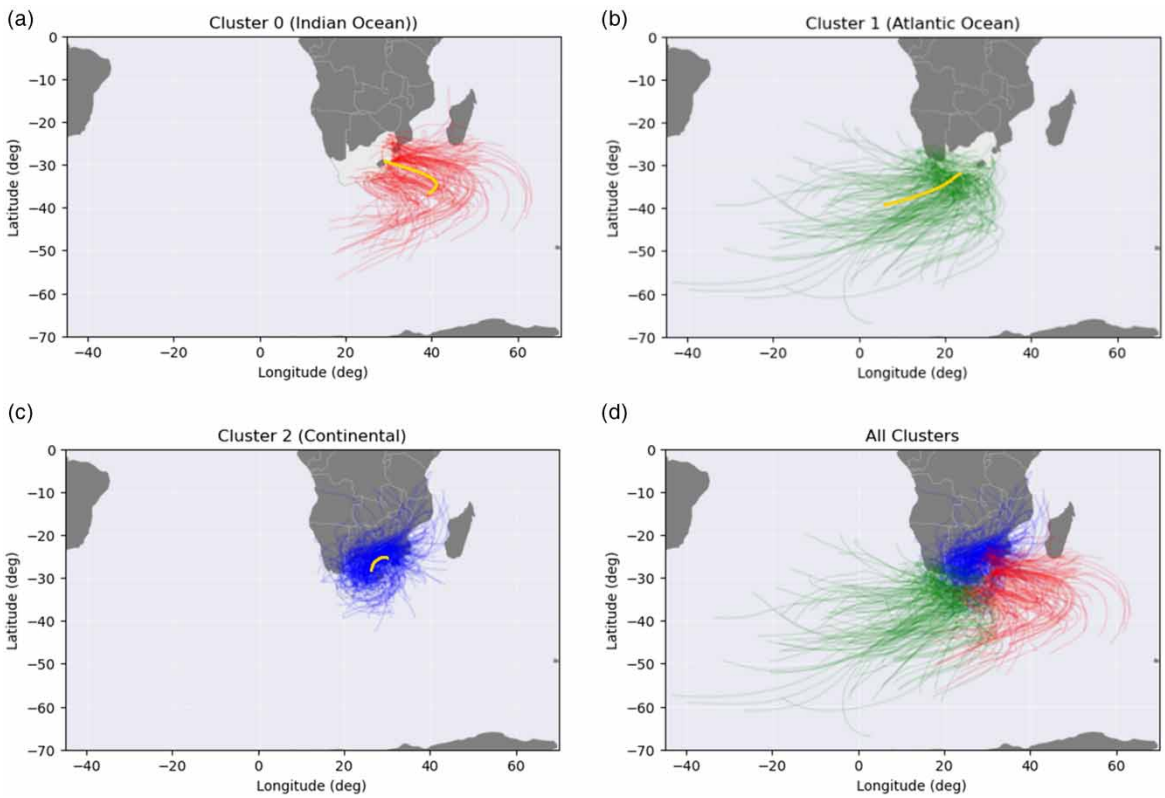


Figure 4 | One in 10 trajectories of (a, top left) cluster 0 (Indian Ocean originating events); (b, top right) cluster 1 (Atlantic Ocean originating events); (c, bottom left) cluster 2 (continentally originating events); and (d, bottom right) all clusters plotted together to show distinct spatial difference.

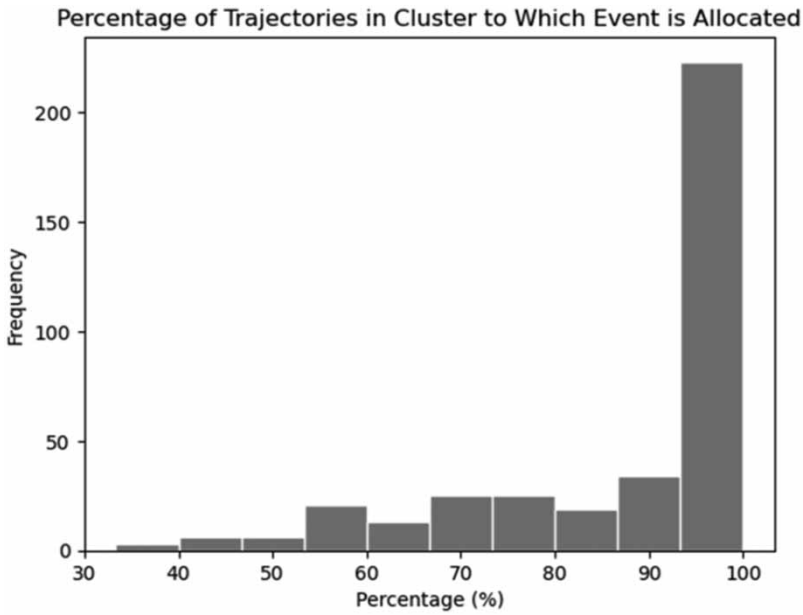


Figure 5 | Event allocation. This chart displays the percentage of trajectories of an event that were allocated to the cluster that the event was attributed to.

Only eight events (2.1%) have been allocated by this method to clusters that were originally allocated less than 50% of the trajectories of the event.

The *K*-means clustering technique has been shown to be efficient at allocating events to clusters. However, as 24 (the number of trajectories generated per event) is both even and divisible by 3 (the number of clusters created), there is a possibility that events were erroneously allocated to more than one cluster. For example, if an event had an equal number of trajectories in two or more clusters (this can occur if eight trajectories are allocated to each of the three clusters from the same event, or if 12 trajectories from a single event are allocated to two of the three clusters, for example). Once again, only few events were not adequately allocated by this method. In total, eight events (2.1%) were found to contain an equal number of maximum trajectories in more than one cluster. Of the events allocated to more than one cluster, three were found to be events that were initially allocated to those clusters with less than 50% of their trajectories assigned. Overall, the clustering process resulted in three visually distinct clusters representing the spatial origin of moisture into which 368 of 376 events (98.1%) can be allocated based on the allocation of the greatest number of their trajectories.

The allocation of each of the rainfall events to a cluster is synonymous with allocating an event to a particular causal weather system. This allows for a study of the spatial trends of the extreme rainfall events considered, and the dominance of each cluster as an extreme rainfall-generating mechanism has been determined, as detailed in Table 1.

Table 1 shows that for all clusters, approximately the same percentage of trajectories are allocated to each cluster as events (19 and 18% for cluster 0, 26 and 25% for cluster 1, 55 and 58% for cluster 2, respectively). This indicates that both the clustering process and the event allocation process are efficient and provide further confidence that the clusters are accurate

Table 1 | Allocation of events to clusters

Cluster	Number of trajectories clustered (%[a])	Number of events represented (%[b])	Number of events allocated (%[c])
0 – Indian Ocean	1,663 (19)	123 (33)	64 (18)
1 – Atlantic Ocean	2,319 (26)	170 (45)	91(25)
2 – Continental	4,952 (55)	278 (74)	213 (58)

[a]All percentages of the total rounded to the nearest whole number.
[b]Percentages in this column will not add up to 100% as each event can be represented in one, two, or all three clusters.
[c]Not including the eight events that could not be allocated.

representations of the true extreme rainfall-generating processes. If one or more clusters were allocated a significantly higher percentage of trajectories than events, this would indicate that the cluster contained a small number of trajectories from each of a larger number of events. This would likely not represent a rainfall-generating process but an amalgamation of the trajectories from two or more generating processes and may be an indication that a non-optimal K value had been used when clustering. Further evidence of the efficiency of the clustering process can be attributed to the fact that the variation between the number of events represented and the number of events allocated remains consistent across all clusters.

Table 1 demonstrates that the continentally originating moisture dominates extreme rainfall events with 58% of recorded events being attributed, followed by the Atlantic Ocean contributing 25% and the Indian Ocean contributing 18%. The differences in the spatial origin of these events are stark and provide evidence that extreme rainfall events in South Africa can have very different origins.

4.3. Variation in event magnitude between clusters

The allocation of events to clusters enables analysis of the distribution of rainfall magnitude within and between clusters. Figure 6 shows box-plots of event magnitude within each of the three clusters and illustrates that while the median average extreme rainfall magnitudes for clusters 1 (Atlantic Ocean, 149.3 mm) and 2 (Continental, 150 mm) are approximately equal, the median average for cluster 0 (Indian Ocean, 279 mm) is significantly larger.

Furthermore, the interquartile range of cluster 1 (Atlantic Ocean) rainfall is greater than cluster 2 (continental), indicating marginally greater variation in rainfall magnitude. However, when considering outliers (defined as values with a magnitude greater than the 75% percentile $+ 1.5 \times$ interquartile range), it is cluster 2 (continental) that has produced the largest rainfall event on record.

Table 1 and Figure 6 indicate that of the events considered, cluster 0 (Indian Ocean) events generally contribute greater levels of precipitation during rainfall events, but only represent 18% of all recorded events, whereas the remaining events attributed to cluster 1 (Atlantic Ocean) and cluster 2 (continental) average far lower levels of precipitation but represent the majority of events.

4.4. Spatial distribution of clusters

Figure 7(a)–7(d) show the spatial distribution of the events of each cluster which visually matches the trajectory clusters shown in Figure 4(a)–4(d). As expected, the events that are attributed to Indian Ocean influences are predominantly located along the eastern coastal regions and the events attributed to Atlantic Ocean influences are primarily located along the west

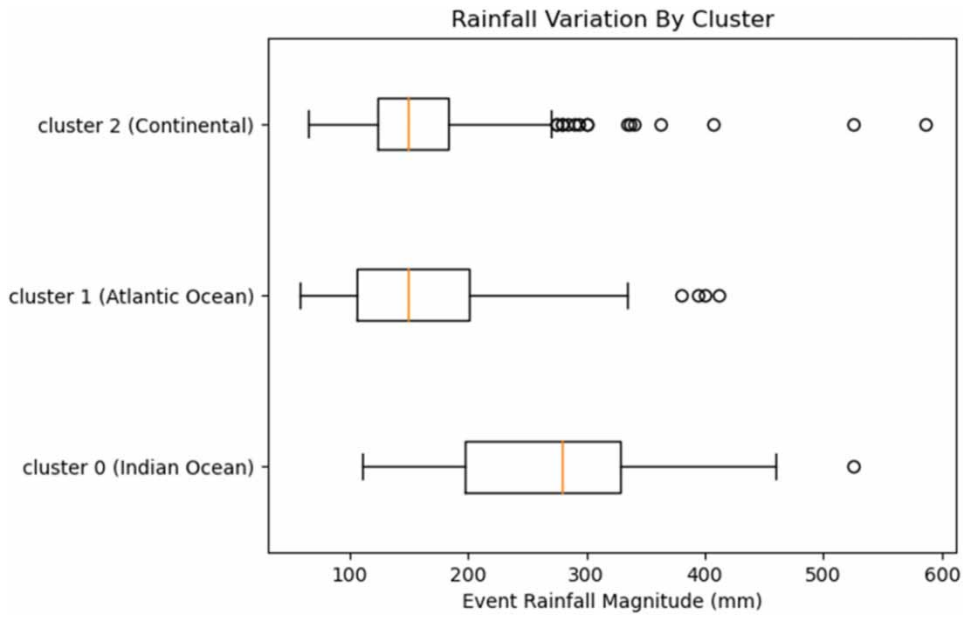


Figure 6 | Rainfall magnitude distribution for historical AMAX events by cluster.

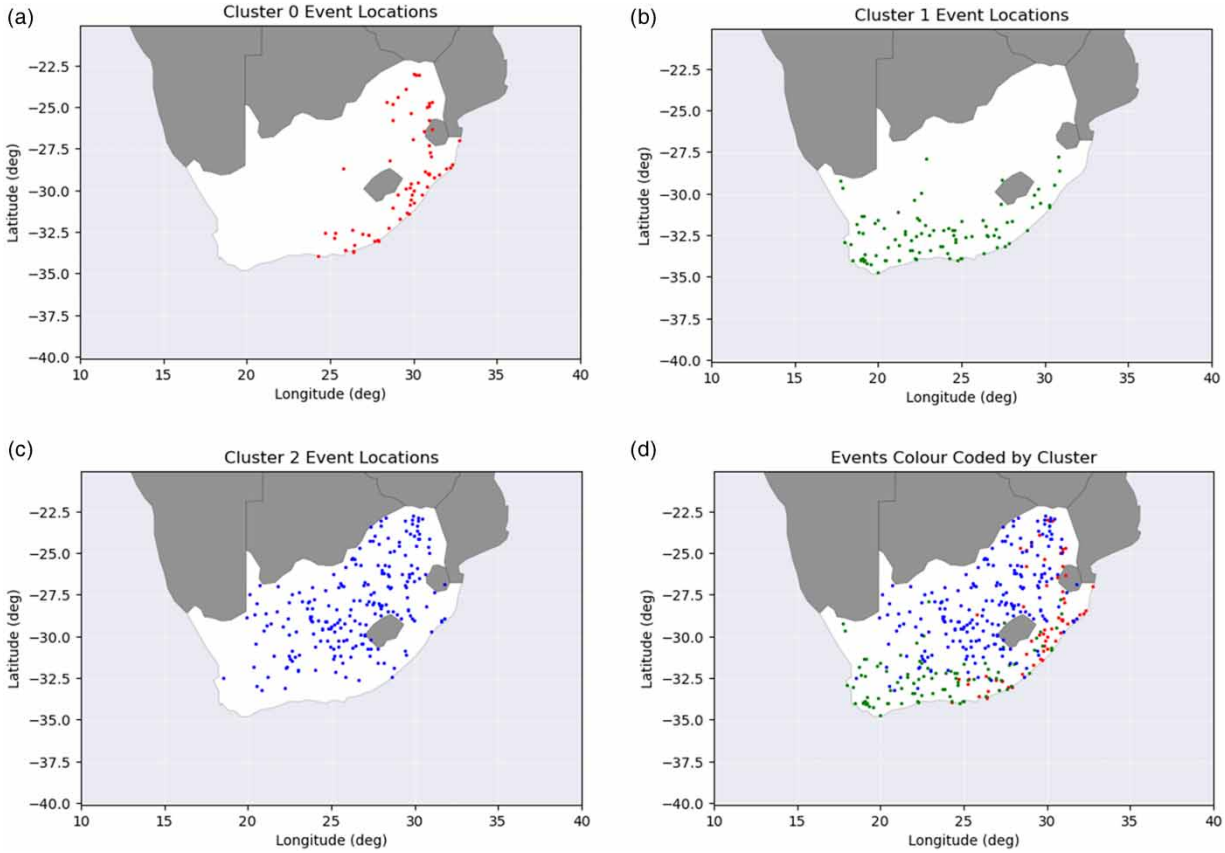


Figure 7 | Event locations for (a, top left) cluster 0 (Indian Ocean originating events); (b, top right) cluster 1 (Atlantic Ocean originating events); (c, bottom left) cluster 2 (continentally originating events); and (d, bottom right) all event locations shown on a map of South Africa.

and southern coastal areas. The events attributed to continental recirculation are found almost exclusively in the interior region.

Figure 7(d) suggests that South Africa can be split roughly into three regions when considering the origin of moisture causing extreme rainfall events. The interior is dominated by continental recirculation, and the coasts are dominated by oceanic influences as would be expected. Furthermore, when considering $K = 3$, the coastline can be split at approximately 25° longitude – the centre of the country – to delineate between coastal regions influenced by the Atlantic Ocean to the west and the Indian Ocean to the east. Whilst this is not an exact boundary, it does provide a rough guide to the origins of extreme rainfall in different regions.

When also considering that the oceanic originating events are confined to the coastal regions, at the base of the plateau that delineates central South Africa, Figure 7(d) indicates that the dominant cause of extreme rainfall in a region of South Africa can be effectively determined by two factors: whether the location is east or west of 25° longitude and whether the location sits on the plateau in the interior or on the sides of, or at the base of, the escarpment that delineates the plateau. Given the large area of the plateau in which rainfall is dominated by continental systems compared to the coastal areas in which oceanic influences dominate it is unsurprising that Table 1 demonstrates that the majority of extreme rainfall events considered are continental in origin.

5. CONCLUSIONS

This study sought to investigate the origins and pathways of atmospheric moisture causing extreme rainfall events in South Africa. To achieve this aim, a process was developed for extracting storm trajectories using data from the HYSPLIT model and clustering these trajectories using unsupervised machine learning techniques. A spatially representative database of observed historical maximum magnitude rainfall events has been considered combined with meteorological data supplied

by the National Oceanic and Atmospheric Administration (NOAA) Air Resources Laboratory NCEP/NCAR Reanalysis 1 project.

The new clustering process developed in this study has revealed three distinct regions in which moisture originates when considering extreme rainfall events in South Africa – South Atlantic Ocean, South Indian Ocean and continentally – and that there are clear differences in the spatial and temporal distributions of these events. The coastal regions of South Africa are predominantly influenced by the respective adjacent oceans with cluster 0 events originating from the South Indian Ocean and dominating the east coast, whereas the west coast is predominantly influenced by moisture originating in the South Atlantic Ocean (cluster 1). Furthermore, the central region of South Africa is dominated by continentally originating moisture. Clear differences in rainfall magnitude have been identified with cluster 0 (Indian Ocean) accounting for 18% of events with an average magnitude of 279 mm, cluster 1 (Atlantic Ocean) accounting for 25% of events with an average magnitude of 149.3 mm, and cluster 2 (continental) accounting for the majority of events (58%) with an average magnitude of 150 mm. When considering South Africa as a whole, it appears as though the least frequent events are also the ones that carry the largest magnitude; however, the clustering process has identified that these events are predominantly found in the eastern region, indicating that it is more appropriate to consider extreme rainfall on a regional and local level when designing infrastructure.

The regions of influence of the three regions of moisture origin have been found to be demarcated approximately by the line of 25° longitude and the escarpment that delineates the central plateau. The approximate demarcation at 25° longitude is most likely due to this being approximately the boundary between the eastern arid zones (influenced by cluster 1, Atlantic Ocean) and the western temperate zones (influenced by cluster 0, Indian Ocean). Cluster 2 (continental) events are spread approximately evenly across the temperate and arid zones due to this cluster likely being dominated by the altitude of the plateau as a causal rainfall mechanism, rather than oceanic influences, and the approximately even area of the central plateau in both the temperate and arid regions.

This study has demonstrated that the combination of the HYSPLIT model and unsupervised clustering techniques is capable of developing insights into the spatial origin, dominance, and magnitude of extreme rainfall. These findings opens-up the possibility for further studies into how climate change might affect the processes governing extreme rainfall in the region and the potential consequences for the safety of critical infrastructure.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model. The five anonymous reviewers of the manuscript all provided helpful feedback on earlier versions of the manuscript, and their contributions were acknowledged.

STATEMENTS AND DECLARATIONS

The authors have no relevant financial or non-financial interests to disclose

AUTHOR CONTRIBUTIONS

Conceptualization: Thomas Rodding Kjeldsen and Rhys Phillips; Data curation: Katelyn Johnson; Methodology: Thomas Kjeldsen, Andrew Barnes, Katelyn Johnson, and Rhys Phillips; Formal analysis and investigation: Rhys Phillips; Writing – original draft preparation: Rhys Phillips; Writing – review and editing: Rhys Phillips, Thomas Rodding Kjeldsen, Andrew Barnes, and Katelyn Johnson.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahern, M., Kovats, R. S., Wilkinson, P., Few, R. & Matthies, F. 2005 Global health impacts of floods: Epidemiologic evidence. *Epidemiologic Reviews* **27**, 36–46. https://doi.org/10.1093/epirev/mxi004.

AON 2020 *Weather, Climate & Catastrophe Insight*. Available from: http://catastropheinsight.aon.com.

Ashrafi, K., Shafiepour-Motlagh, M., Aslemand, A. & Ghader, S. 2014 Dust storm simulation over Iran using HYSPLIT. *Journal of Environmental Health Science and Engineering* **12**, 1–9.

Barnes, A. P., McCullen, N. & Kjeldsen, T. R. 2019 The atmospheric origins of extreme rainfall in the UK. In: *Proceedings of the IMA's 4th International Flood Risk Conference*. Available from: https://researchportal.bath.ac.uk/en/publications/atmospheric-origins-of-extreme-rainfall-in-the-uk.

Barnes, A. P., Santos, M. S., Garijo, C., Mediero, L., Prosdocimi, I., McCullen, N. & Kjeldsen, T. R. 2020 Identifying the origins of extreme rainfall using storm track classification. *Journal of Hydroinformatics* **22** (2), 296–309. https://doi.org/10.2166/hydro.2019.164.

Cui, L., Song, X. & Zhong, G. 2021 Comparative analysis of three methods for HYSPLIT atmospheric trajectories clustering. *Atmosphere* **12** (6), 698.

de Waal, J. H., Chapman, A. & Kemp, J. 2017 Extreme1-day rainfall distributions: Analysing change in the Western Cape. *South African Journal of Science* **113** (7/8), 1–8. https://doi.org/10.17159/sajs.2017/20160301.

di Baldassarre, G., Montanari, A., Lins, H., Koutsoyiannis, D., Brandimarte, L. & Blschl, G. 2010 Flood fatalities in Africa: From diagnosis to mitigation. *Geophysical Research Letters* **37** (22). https://doi.org/10.1029/2010GL045467.

du Plessis, J. A. & Burger, G. J. 2015 Investigation into increasing short-duration rainfall intensities in South Africa. *Water SA* **41** (3), 416–424. https://doi.org/10.4314/wsa.v41i3.14.

Evans, B. P., Xue, B. & Zhang, M. 2019 What's inside the black-box? A genetic programming method for interpreting complex machine learning models. In *GECCO 2019 - Proceedings of the 2019 Genetic and Evolutionary Computation Conference*, pp. 1012–1020. https://doi.org/10.1145/3321707.3321726.

Hahn, A., Schefuß, E., Andò, S., Cawthra, H. C., Frenzel, P., Kugel, M., Meschner, S., Mollenhauer, G. & Zabel, M. 2017 Southern Hemisphere anticyclonic circulation drives oceanic and climatic conditions in late Holocene southernmost Africa. *Climate of the Past* **13** (6), 649–665. https://doi.org/10.5194/cp-13-649-2017.

Hartigan, J. A. & Wong, M. A. 1979 Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28** (1), 100–108.

Hurst, T. & Davis, C. 2017 Forecasting volcanic ash deposition using HYSPLIT. *Journal of Applied Volcanology* **6** (1), 1–8.

International Food Policy Research Institute (IFPRI) 2014 *Atlas of African Agriculture Research and Development*. International Food Policy Research Institute (IFPRI), Washington D.C.

Johnson, K. A. & Smithers, J. C. 2020 Updating the estimation of 1-day probable maximum precipitation in South Africa. *Journal of Hydrology: Regional Studies* **32**. https://doi.org/10.1016/j.ejrh.2020.100736.

Jury, M. R. 2015 Passive suppression of South African rainfall by the Agulhas Current. *Earth Interactions* **19** (13). https://doi.org/10.1175/EI-D-15-0017.1.

Karozis, S., Sfetsos, A., Gounaris, N. & Vlachogiannis, D. 2021 An assessment of climate change impact on air masses arriving in Athens, Greece. *Theoretical and Applied Climatology* **145** (1–2), 501–517.

Khandlhela, M. & May, J. 2006 Poverty, vulnerability and the impact of flooding in the Limpopo Province, South Africa. *Handbook of Environmental Chemistry, Volume 5: Water Pollution* **39** (2), 275–287. https://doi.org/10.1007/s11069-006-0028-4.

le Maitre, D., Kotzee, I., le Roux, A. & Ludick, C. 2019 *Floods Current State and Implications of Climate Change*. Available from: https://pta-gis-2-web1.csir.co.za/portal/apps/GBCascade/index.html?appid=33d9a846cf104e1ea86ba1fa3d197cbd [Accessed 24 September 2021].

Lumbroso, D. 2020 Flood risk management in Africa. *Journal of Flood Risk Management* **13** (3), 1–5. https://doi.org/10.1111/jfr3.12612.

Ma, Y., Wang, M., Wang, S., Wang, Y., Feng, L. & Wu, K. 2021 Air pollutant emission characteristics and HYSPLIT model analysis during heating period in Shenyang, China. *Environmental Monitoring and Assessment* **193**, 1–14.

Mabuse, N. 2021 *Officials Say 40 Killed in South African Floods; More Rain Predicted*. Available from: http://edition.cnn.com/2011/WORLD/africa/01/18/south.africa.floods/ [Accessed 18 March 2021].

McGowan, H. & Clark, A. 2008 Identification of dust transport pathways from Lake Eyre, Australia using HYSPLIT. *Atmospheric Environment* **42** (29), 6915–6925.

Odoulami, R. C., Wolski, P. & New, M. 2020 A SOM-based analysis of the drivers of the 2015–2017 Western Cape drought in South Africa. *International Journal of Climatology* **41**, 1518–1530. https://doi.org/10.1002/joc.6785.

Roffe, S. J., Fitchett, J. M. & Curtis, C. J. 2019 Classifying and mapping rainfall seasonality in South Africa: A review. *South African Geographical Journal* **101** (2), 158–174. https://doi.org/10.1080/03736245.2019.1573151.

Rousseeuw, P. J. 1987 Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20** (C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Santos, M. S., Mediero, L., Lima, C. H. R. & Moura, L. Z. 2018 Links between different classes of storm tracks and the flood trends in Spain. *Journal of Hydrology* **567**, 71–85. https://doi.org/10.1016/j.jhydrol.2018.10.003.

Schulze, R. E. & Schütte, S. 2019 *Update of Potential Climate Change Impacts on Relevant Water Resources Related Issues in the UMgeni and Surrounding Catchments Using Outputs From Recent Global Climate Models as Inputs to Appropriate Hydrological Models*. Centre for Water Resources Research, Pietermaritzburg.

SciKit Learn 2020 *Silhouette Coefficient*. Available from: https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient [Accessed 8 January 2021].

South African Weather Service. 2019 *Trends in Extreme Climate Indices in South Africa*. Available from: https://www.weathersa.co.za/Documents/Corporate/WMOExtremeClimateIndicesreport2019.pdf [Accessed 19 November 2020].

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D. & Ngan, F. 2015 Noaa's HYSPLIT atmospheric transport and dispersion modeling system. *Bulletin of the American Meteorological Society* **96** (12), 2059–2077. https://doi.org/10.1175/BAMS-D-14-00110.1.

Tan, X., Gan, T. Y. & Chen, Y. D. 2018 Moisture sources and pathways associated with the spatial variability of seasonal extreme precipitation over Canada. *Climate Dynamics* **50** (1–2), 629–640. https://doi.org/10.1007/s00382-017-3630-0.

UNOOSA 2019 *UNOOSA Activates International Charter for Floods and Mudslides in South Africa*. Available from: https://disasterscharter.org/web/guest/activations/-/article/flood-insouth-africa-activation-605- [Accessed 18 March 2021].

Wallace, J. M. & Hobbs, P. v., 2006 In: *Atmospheric Science: An Introductory Survey* (Hele, J. ed.). Academic Press. Available from: https://books.google.co.uk/books?hl=en&lr=&id=HZ2wNtDOU0oC&oi=fnd&pg=PP1&ots=C5LIlgm-S0&sig=-uLbFWUkWOB8p1PRnGWuaZPUPuA&redir_esc=y#v=onepage&q&f=false.

Wang, S., Zhang, W. & Chen, F. 2019 Simulation of drainage capacity in a coastal nuclear power plant under extreme rainfall and tropical storm. *Sustainability* **11** (3), 642.

Winsemius, H. C., Aerts, J. C. J. H., van Beek, L. P. H., Bierkens, M. F. P., Bouwman, A., Jongman, B., Kwadijk, J. C. J., Ligtvoet, W., Lucas, P. L., van Vuuren, D. P. & Ward, P. J. 2016 Global drivers of future river flood risk. *Nature Climate Change* **6** (4), 381–385. https://doi.org/10.1038/nclimate2893.

World Bank 2021 *South Africa: Urbanization From 2010 to 2020*. Available from: https://www.statista.com/statistics/455931/urbanization-in-south-africa/ [Accessed 24 September 2021].

Xia, Z., Butorovic, N. & Yu, Z. 2020 The influence of synoptic weather types and moisture transport pathways on precipitation isotopes in Southern Patagonia. *Atmosphere* **11** (5), 514.

Ziervogel, G., New, M., Archer van Garderen, E., Midgley, G., Taylor, A., Hamann, R., Stuart-Hill, S., Myers, J. & Warburton, M. 2014 Climate change impacts and adaptation in South Africa. *Wiley Interdisciplinary Reviews: Climate Change* **5** (5), 605–620. https://doi.org/10.1002/wcc.295.

Zuma, B. M., Luyt, C. D., Chirenda, T. & Tandlich, R. 2012 Flood Disaster Management in South Africa : Legislative framework and current challenges. In *International Conference on Applied Life Sciences (ICALS)*, pp. 127–132.

First received 27 June 2023; accepted in revised form 5 December 2023. Available online 21 December 2023