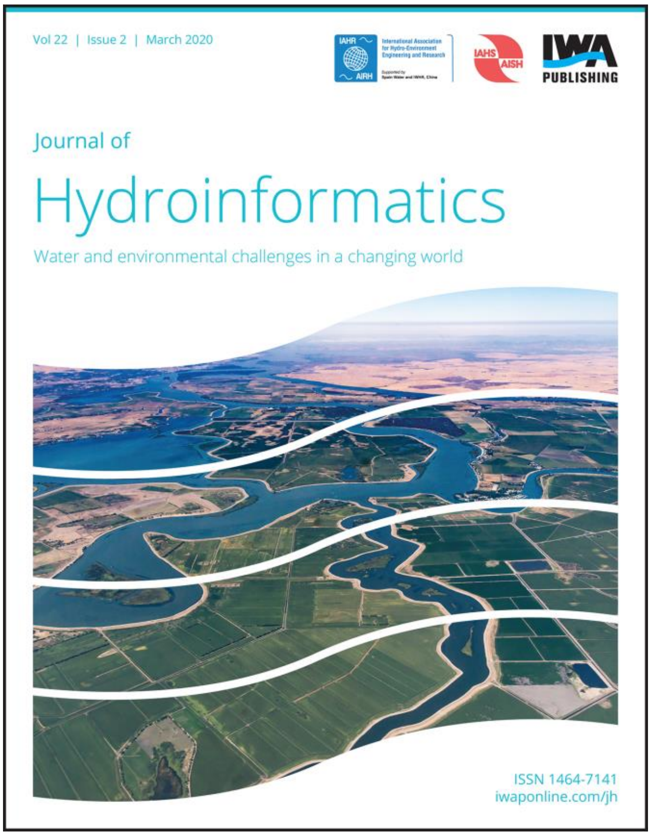

Identifying the origins of extreme rainfall using storm track classification

Andrew Paul Barnes, Marcus Suassuna Santos, Carlos Garijo, Luis Mediero, Ilaria Prosdocimi, Nick McCullen and Thomas Rodding Kjeldsen

ABSTRACT

Identifying patterns in data relating to extreme rainfall is important for classifying and estimating rainfall and flood frequency distributions routinely used in civil engineering design and flood management. This study demonstrates the novel use of several self-organising map (SOM) models to extract the key moisture pathways for extreme rainfall events applied to example data in northern Spain. These models are trained using various subsets of a backwards trajectory data set generated for extreme rainfall events between 1967 and 2016. The results of our analysis show 69.2% of summer rainfall extremes rely on recirculatory moisture pathways concentrated on the Iberian Peninsula, whereas 57% of winter extremes rely on deep-Atlantic pathways to bring moisture from the ocean. These moisture pathways have also shown differences in rainfall magnitude, such as in the summer where peninsular pathways are 8% more likely to deliver the higher magnitude extremes than their Atlantic counterparts.

Key words | extreme rainfall, frequency analysis, HYSPLIT, self-organising maps, trajectory analysis

Andrew Paul Barnes (corresponding author)
Nick McCullen
Thomas Rodding Kjeldsen
Department of Architecture & Civil Engineering,
University of Bath,
BA2 7AY Bath, UK
E-mail: *a.p.barnes@bath.ac.uk*

Marcus Suassuna Santos
Department of Hydrology,
Brazilian Geological Survey – CPRM,
SBN Quadra 2, Asa Norte, 70040-904 Brasilia,
Brazil
and
Department of Civil and Environmental Engineering,
University of Brasilia,
Darcy Ribeiro Campus, 70910-900 Brasilia,
Brazil

Carlos Garijo
Luis Mediero
Department of Civil Engineering: Hydraulics, Energy and Environment, ETSI de Caminos, Canales y Puertos,
Universidad Politécnica de Madrid,
Madrid, Spain

Ilaria Prosdocimi
Department of Environmental Sciences, Informatics and Statistics,
Ca' Foscari University of Venice,
Venice, Italy

INTRODUCTION

Floods generated by extreme weather events continue to be a global issue causing widespread social and economic damage. Effective flood risk management requires estimates of the frequency and magnitude of future flood characteristics, such as, for example, the magnitude of the design rainfall and flood events with a return period of 100 or 10,000 years. Such estimates are obtained through frequency analysis by fitting statistical extreme value distributions directly to past extreme events. Traditional frequency analysis techniques do not account for differences in the underlying processes causing extreme events. The importance of accounting for different event-generating processes in frequency analysis has been discussed in several studies. Waylen & Woo (1982) separated an annual maximum (AMAX) series of flood peak into events caused by rainfall and snowmelt and fitted a mixture distribution consisting of two Gumbel distributions. Hirschboeck (1987) manually divided a flood distribution into eight subcategories each representing floods caused by different atmospheric patterns. This resulted in a set of distributions, each with significantly different structure with some containing multiple peaks, whereas others showed a distinct

doi: 10.2166/hydro.2019.164

generalised extreme value distribution (Hirschboeck 1987). Merz & Blöschl (2003) utilised a process-oriented method by separating the initial flood distribution by generating mechanism in Austrian catchments, which was then used to show 'short rain' floods generally happened in the southern part of the country. Villarini & Smith (2010) found that the upper tail of flood distributions in the eastern part of the US is influenced by tropical cyclones. Kjeldsen *et al.* (2018) studied extreme rainfall in South Korea and used the information published by the Korean Meteorological Administration to create AMAX series of one-day rainfall caused by typhoons and non-typhoons, respectively.

However, while most studies argue that improved process understanding will improve the reliability of model predictions, there is still a need to develop objective methods for distinguishing between events generated by different mechanisms (Kjeldsen *et al.* 2018). In addition, the benefits of process-oriented techniques come at a computational cost; previously, constraints regarding the availability of data and computational power have limited our ability to identify these generation mechanisms. Where data have been available, it has generally been provided in small sample sizes. This limitation also explains the preference for non-process-based methods, which are less computationally expensive (Hirschboeck 1987). Despite this, increasing amounts of data are becoming available such as through the Hybrid Single-Particle Lagrangian Integrated Trajectory Model (HYSPLIT) (Stein *et al.* 2015) which opens up new opportunities to take advantage of auxiliary knowledge regarding these extreme events.

This study will focus on the different processes controlling extreme precipitation events in the Douro catchment, located in north-western Spain. Mediero *et al.* (2014) found a general decrease trend in AMAX flood series in Spain. In addition, a recent study found that floods towards the north of the Iberian Peninsula are trending towards early winter (Blöschl *et al.* 2017). Santos *et al.* (2018) partially explained such decreasing trends by a negative trend in the moisture amount that arrives in Spain, more evident in the case of Continental storms. Without high-resolution knowledge of how the underlying processes effect flood distributions, current models come with higher process uncertainty.

Current literature explores the large-scale atmospheric processes which influence global rainfall variation. Utsumi *et al.* (2016) found that the main driver for precipitation in central Europe is extratropical cyclones. The authors found a similar result in the Mediterranean but with a higher tendency to be manipulated by extratropical cyclones. Further studies have identified the tropical–subtropical North Atlantic corridor (a corridor stretching from the Gulf of Mexico and the Caribbean Sea to the Mediterranean) to influence moisture fluxes on the Iberian Peninsula (Gimeno *et al.* 2010a, 2010b*;* Scoccimarro *et al.* 2018). Jorba *et al.* (2004) extracted the high-level trajectories arriving in the Barcelona area by using HYSPLIT, clustering them into 10 different patterns describing the 2D (latitude/longitude) pathways. Such work highlights the tropospheric circulation patterns that influence the Barcelona area by identifying that the main flows come from the Atlantic, 5,500 m above sea level (a.s.l.). However, this study focussed on the upper portion of the atmosphere and therefore is not directly useful for identifying moisture transport systems, which are generally found in the lower 2,000 m (Wallace *et al.* 1977). More recently, links have been drawn between low-level trajectory classifications and a temporal trend in their occurrence, by using HYSPLIT to extract trajectories associated with flood events and classifying them using a different clustering method (*k*-means). The results of this analysis showed the Continental storms appeared to be more common than their Atlantic/Mediterranean counterparts when aligned with extreme flood timings (Santos *et al.* 2018).

Moisture pathways associated with seasonal extremes have been identified in Canada by following a similar approach, extracting and classifying trajectories for extreme rainfall events at varying altitudes between 0 and 5,000 m (Tan *et al.* 2017). The study identifies nine spatially coherent regions using self-organising maps (SOMs), highlighting the key moisture sources related to the seasonal extreme precipitation. Despite this, there is limited consideration for varying the number of clusters; the algorithm was initialised with and no indication of whether a numerically accurate solution was produced.

Methods for classifying these moisture trajectories can be grouped into two main categories: supervised and unsupervised. Supervised methods for classification rely on having a training data set with both inputs and known outputs. These methods are most useful for identifying similarities and differences between the known classifications.

An example of a supervised trajectory classification algorithm is TraClass (Lee *et al.* 2008), which utilises the known class labels to identify descriptive areas of the trajectories, which can be used to differentiate each class. Unsupervised classification methods do not require a training set with known class labels. Instead, these methods can be used to identify groups of numerically related input vectors. The most popular unsupervised methods are *k*-means and SOMs, which have recently been shown to successfully identify trajectory groups (Owens & Hunter 2000; Lee *et al.* 2008; Tan *et al.* 2017; Santos *et al.* 2018).

This paper aims to use SOMs both to identify the key moisture pathways, which lead to AMAX rainfall and to highlight the magnitude differences between these classifications. First, the selected case study and data used are described. Second, the classification methodology and models are introduced. Third, the results of the classification model development are presented before final conclusions are presented.

METHODS

Precipitation and trajectory data

This analysis focuses on the Douro catchment located in a north-western region of Spain. AMAX series of one-day precipitation were extracted from 310 gauging stations shown in Figure 1. The data sets available for each station vary in length with some containing data from 1948 and others only containing data from 1967 to 2016. In total, 16,534 one-day rainfall events were extracted. The AMAX data were normalised between 0 and 1 using the equation below, here rainfall$_i$ refers to the AMAX series vector for a given station i:

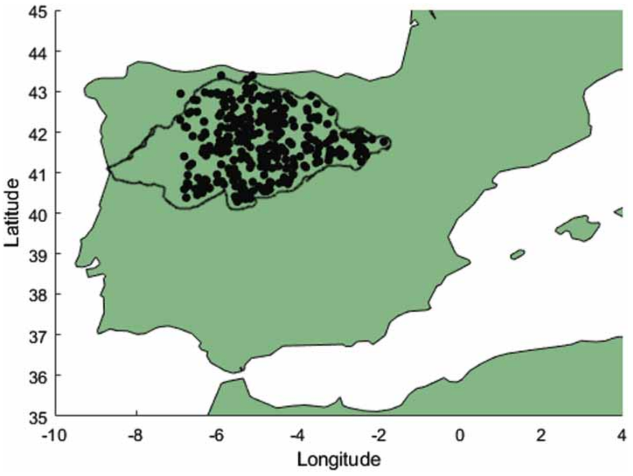

Figure 1 | Locations of gauging stations used to extract AMAX rainfall events and black line indicates the catchment boundary for the Douro river.

$$\text{Normalised rainfall}_i = \frac{\text{rainfall}_i - \min(\text{rainfall}_i)}{\max(\text{rainfall}_i) - \min(\text{rainfall}_i)}$$

This provides a more comparable view of the rainfall variation between stations, as rainfall magnitudes depend on the height and spatial location, among other variables. Consequently, this normalisation removes differences between stations with high and low rainfall magnitudes, providing a baseline to analyse any magnitude changes without the need to use the station as a dependent variable. The normalisation smooths the empirical cumulative distribution curve of rainfall magnitudes (Figure 2).

For each AMAX event, 24 backwards trajectories were generated using the HYSPLIT system (Stein *et al.* 2015). In order to generate these trajectories, HYSPLIT must be initialised with a start location, start time and run-time, which are then used alongside meteorological files to generate a trajectories. We used NCEP/NCAR reanalysis data files (Kalnay *et al.* 1996) as input for the HYSPLIT model. The primary variables used are related to air pressure, velocity, specific humidity and temperature (PRSS, T02M, U10M, V10M, TPP6, HGTS, TEMP, UWND, VWND, WWND and RELH). For a full description of the variables, refer to NOAA (2003). HYSPLIT uses these files to estimate the storm tracks and water budgets for each storm event. The method used by HYSPLIT is a hybrid between a Lagrangian approach and a Eulerian approach, which allows the relative calculation of the advection, diffusion and particle concentrations (Draxler & Hess 1997; Draxler & Hess 1998; Draxler 1999; Stohl & James 2004).

These trajectories were initiated using a set of combinations of altitudes (10, 410, 810, 1,210, 1,610 and 2,010 m a.s.l.) and times (00:00, 06:00, 12:00, and 18:00) on the day of the event's occurrence. The altitudes were selected to coincide with the expectation of moisture pathways generally existing in the lower 2,000 m of the atmosphere (Wallace *et al.* 1977). The length of the

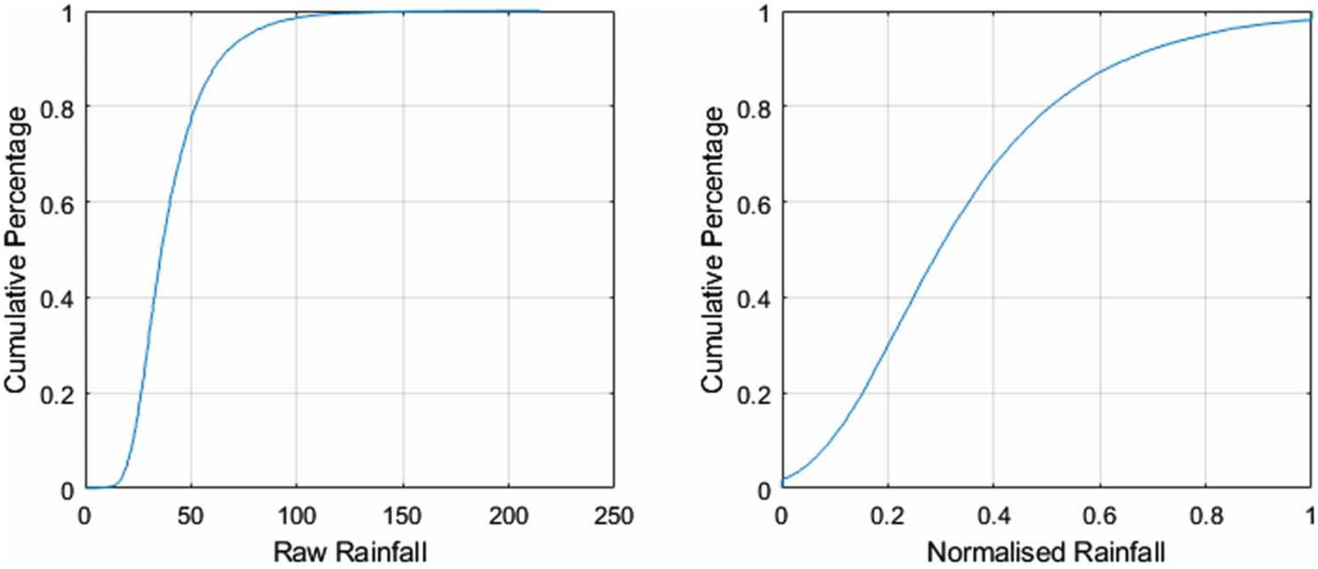

Figure 2 | Empirical cumulative distribution for all rainfall data both before normalisation (left) and after (right).

backwards trajectories was fixed to 48 h before the given initiation time and resulted in a total of 331,728 successfully extracted storm tracks, five examples of the output from HYSPLIT are presented in Figure 3. Each of these trajectories consists of 49 points identifying the position of the air parcel at each hourly interval. Each point has the following information associated with it: latitude, longitude, altitude, specific humidity and atmospheric pressure. For the purposes of this study only latitude, longitude and altitude are used, as the goal is to identify the spatial origin of these events.

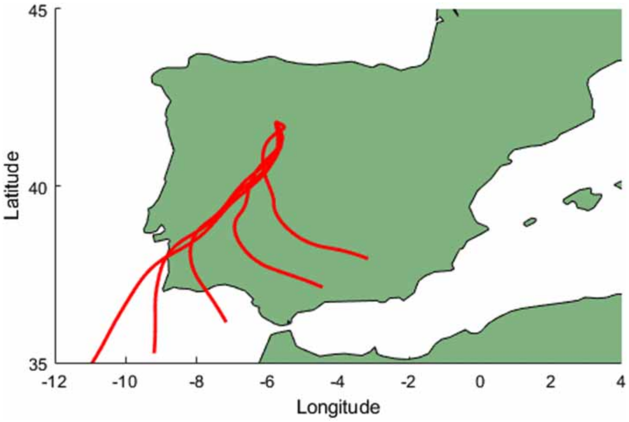

Figure 3 | Five backwards trajectories from a single AMAX event starting from the northernmost point.

Trajectory classification

The SOM approach was adopted to classify these trajectories, due to a significant number of successful hydrological applications (Kalteh *et al.* 2008; Fahimi Yaseen & El-shafie 2016; Tan *et al.* 2017). A self-organised map is a type of neural network architecture used for classifying or reducing the dimensions on input data. It does this through the unsupervised learning of a data set to produce its discretised representation, which is often referred to as a 'map'. The purpose of this is to describe the relationships between the clusters. An example of a SOM is given in Figure 4. SOMs create this map through competitive learning where each output (or class) competes to represent a given input vector (Kohonen & Honkela 2007). For each input vector, the closest node is selected and moved closer to the given input vector; a neighbourhood function is then applied such that the neighbours of the node are also moved although to a lesser extent. This then trickles through the network until the movement is next to none. For a full and detailed description of how the SOM method works, see Kohonen & Honkela 2007.

The key benefit of SOMs is the assurance that inputs, which are close in the original high-dimensional input space, are close in the classified low-dimensional space (output), which is not guaranteed by other procedures such as

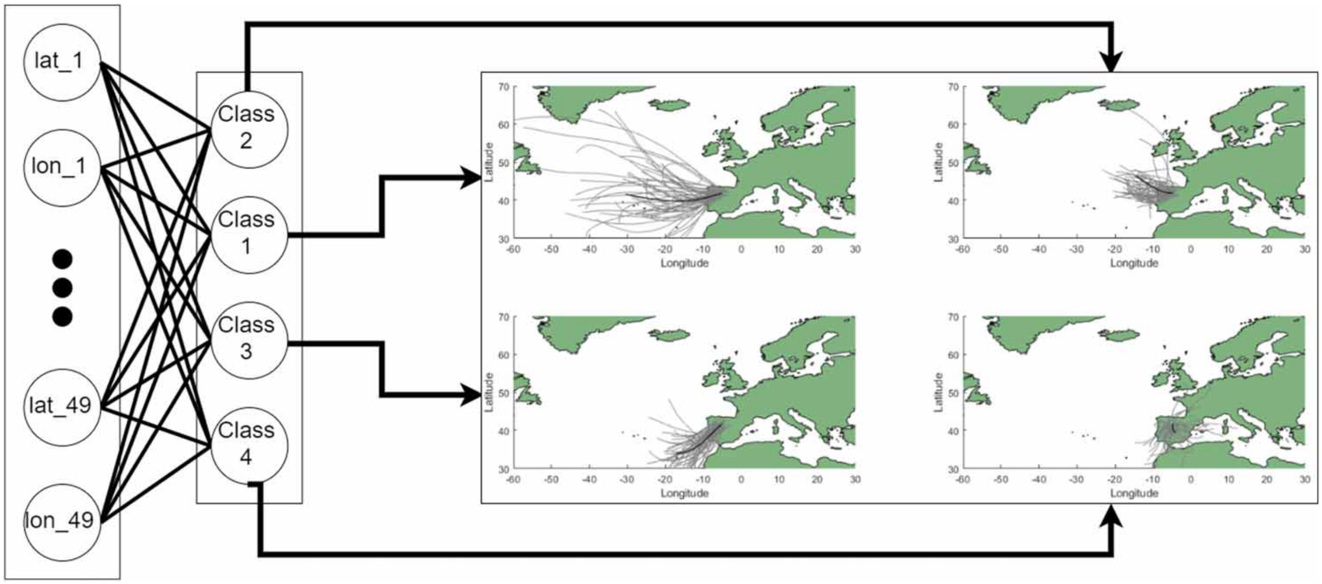

Figure 4 | A sample SOM architecture with 98 input nodes representing the 49 latitude and longitude points of a trajectory. All input nodes are connected via weighted edges to the four output nodes (classifications). These output nodes are arranged into a neighbourhood grid before training which results in classifications being similar to their neighbours. For example, Classes 2 and 3 contain trajectories which are closer to Class 1 than those contained in Class 4.

k-means. This works through for each training item (or in our case trajectory) updating both the closest matching output node and its neighbouring nodes. Further to this, although the SOM approach requires a determined number of output nodes or clusters, it does not require an assumption on the distribution of the data such as would be required by, for example, a Gaussian mixture method (Zhuang *et al.* 1996).

Classifier selection

A classification model accepts many inputs and reduces them into a single class. Common approaches to classifying trajectories rely on the development of a single classifier, such as in Tan *et al.* (2017) who trained a single model and did not consider varying the SOM parameters such as map size and data sampling. However, such approaches fail to capture the relationship between variables in the data set and the classifications. Our approach will explore the differences which occur when varying two key parameters, the subset of data used for training and the size of the map. Further to this, the separation of the data set into two training sets for distinct summer/winter classifiers is used to aid in the identification of seasonal variation in the classifications generated. Here, we define summer and winter to cover the warm (May–September) and cool (October–April) seasonal variations, following a similar pattern to that used in the previous work (Tan *et al.* 2017). Table 1 describes the training data set for the four classifiers

Table 1 | Characteristics of the five classifiers used: here, date filtering is inclusive

Classifier	Code name	No. of classes	Altitude	Date filtering	No. of inputs	No. of trajectories
Primary-4	PR-4	4	No	None	98	331,728
Primary-9	PR-9	9	No	None	98	331,728
3D	3D-9	9	Yes	None	147	331,728
Summer	SUM-9	9	No	May–September	98	88,776
Winter	WIN-9	9	No	October–April	98	242,952

The number of inputs is defined as the size of each input vector before it is classified, and the number of classes is the number of output nodes available to the SOM.

chosen and identifies which parts of the input vectors are included as well as any filtering on the trajectories.

Classifier optimisation

To ensure a numerically stable solution is reached each classifier is first trained with 10 different map sizes. A map size refers to the number of nodes in a square grid, for example, a 2 × 2 map size would result in four output classifications, and a 3 × 3 map size results in nine output classifications. These maps indicate how the output nodes (clusters) are arranged. The example given in Figure 4 shows four clusters arranged in a 2 × 2 map with Class 1 having neighbours 2 and 3, Class 3 having neighbours 1 and 4, etc. An alternative approach would be to use a 4 × 1 map. This structure enables the SOM architecture to identify clusters which may be related, such as Classes 1 and 2 from Figure 4 are much more similar than Classes 1 and 4. A batch processing approach is used for training because it has been shown to be an order of magnitude faster than the alternative linear training approach (Kohonen & Honkela 2007). In addition, two error metrics are calculated for each classifier. The quantisation error calculates the root-mean-squared Euclidean distance between each training sample and its best matching node (BMN). The BMN is defined as the output node with the highest level of activation when a given input is used. The topological error calculates the percentage of training samples which have a first and second BMN which are not adjacent on the output map, ensuring topological consistency (Kiviluoto 1996).

Figure 5 shows the errors produced for each of the four classifiers. Each graph gives the squared dimension (map size) of each model on the *x*-axis, such that a value of two equates to a 2 × 2 output map, which contains four classifications. As is expected in any clustering procedure, the quantisation error decreases with the increase in the number of potential classifications as illustrated in all three quantisation error graphs. Despite this, there is still a clear indication of the numerically superior classifiers as the magnitude of the errors varies significantly.

The large difference in quantisation error could be the curse of dimensionality, an umbrella term for the disadvantages caused by having large input vectors which do not occur when using lower-dimensional vectors (Kohonen & Honkela 2007). The 3D classifier uses 147 input variables, whereas the other only uses 98, and it is known that

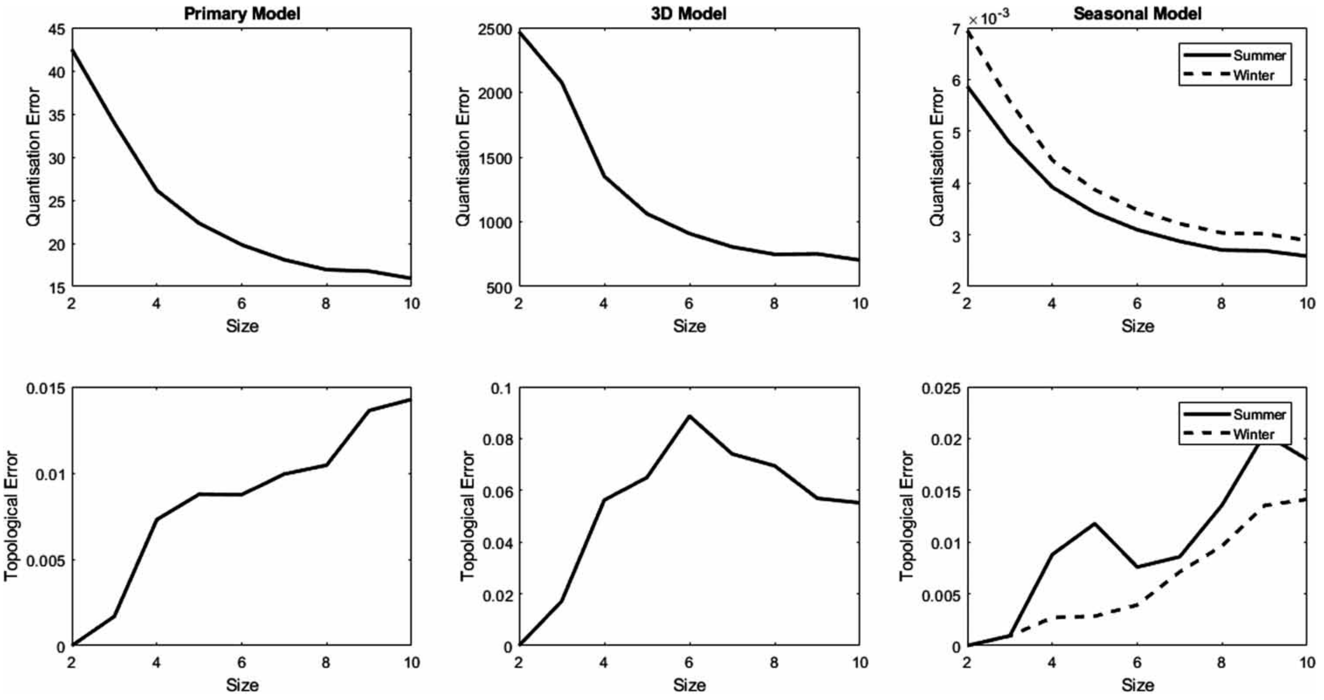

Figure 5 | The quantisation (top) and topological (bottom) errors for each of the three classifiers.

increasing the number of input variables in already high-dimensional problems can cause a decrease in search performance such as when selecting the BMN (Marimont & Shapiro 1979). These differences are not replicated in the topological errors, which tend to increase with an increase in map size. For instance, the Primary classifier shows a consistently lower error indicating a better topological fit of the data. Due to the increasing nature of the topological error, it is concluded that the smaller maps are more numerically accurate. Therefore, this study will utilise a 2×2 map and a 3×3 map for the primary and both seasonal models to present any differences between having either a high topological or quantisation error.

RESULTS

This section discusses the resulting classifications for each of the five models presented in Table 1. It highlights the key moisture pathways as well as their prominence in the data set. This section concludes on an analysis of the rainfall magnitudes for each cluster.

Classifications

Beginning with the Primary classifiers, Figure 6 shows the classifications generated in the SOM with a 2×2 map size, which is referred to as the Primary 4-SOM classifier (PR-4) indicating it has four output classifications. The results show four patterns: Class 1 contains a mid-/western Atlantic originating path, Class 2 includes those coming from the north Atlantic, Class 3 contains the storm tracks coming from the south Atlantic, and Class 4 contains the continental or recirculation pathways. The proportion of trajectories falling into each classification is provided in Table 2. The two dominant pathways are the recirculation and mid-/western Atlantic classes which contain 46.5% (Class 4) and 35.2% (Class 1) of the sample trajectories, respectively. The least common class was the north Atlantic class (Class 2), which only contained 6.7% of the sample; however, this was only slightly lower than the 11.6% of storm tracks classified as south Atlantic storms (Class 3).

Comparing the results of PR-4 with those of a larger map size reaffirms the key pathways identified. Figure 7 shows the classifications produced from the SOM with a 3×3

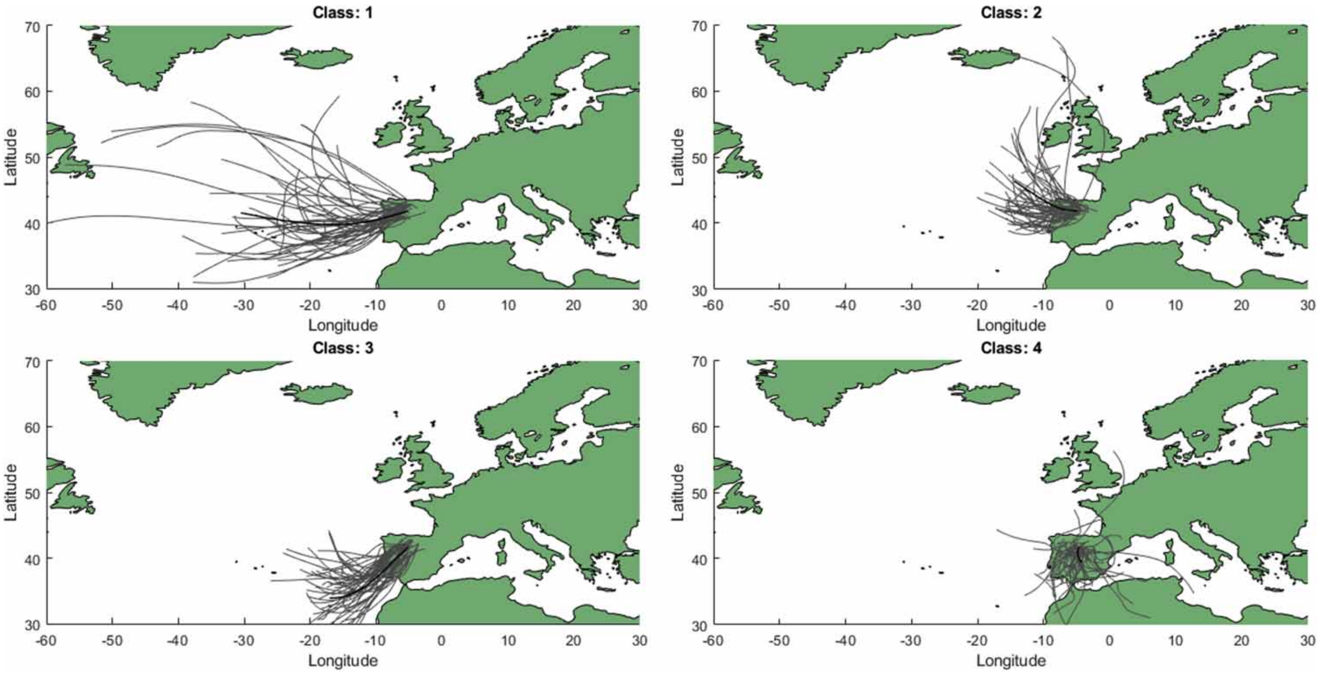

Figure 6 | A sample of trajectories for each cluster in the Primary 4-SOM classifier (PR-4). The classifications plot shows 50 randomly selected trajectories for each class (coloured grey) with a mean trajectory across all relevant trajectories coloured in black. The proportion of trajectories within each class is given in Table 2.

Table 2 | Proportion of trajectories which are classified under each of the four classes within PR-4

	Class			
	1	**2**	**3**	**4**
%	35.2	6.7	11.6	46.5

map size in the Primary classifier (PR-9). Here, the classifications appear to be more refined versions of the ones present in PR-4. For example, Classes 7, 4, 1 and 5 appear to be subclasses of the mid-/western Atlantic class (Class 1) from PR-4. The same can be seen with the continental class (Class 4) in PR-4, which is broken down into Classes 3 and 6 in the PR-9. Moreover, the north Atlantic classification (Class 2) from PR-4 appears unchanged in the PR-9 (again Class 2). The proportions for each class in PR-9 are expectedly lower due to the greater spread amongst the range of classes as shown in Table 3. Despite this, the same pattern appears with continental and mid-/western Atlantic classes (Classes 1 and 4) containing the higher proportions of the sample.

The 3D classifier (3D-9) tended to have the highest numerical errors in comparison to the other classifiers, and on inspection, the tracks appear to have a higher visual variance. The two prominent classes, as shown in Figure 8, such as Classes 1 and 3 differ only in general length and cannot be linked back to classifications in PR-4 or PR-9. However, the spread of trajectories highlighted in Table 4 shows that there is a preference for Classes 1 and 3.

Regarding the seasonal classifiers, they produced the lowest errors during the training phase with the winter classifier producing errors slightly higher than the summer classifier. One possible reason for this disparity could be the quantity of the samples used. Winter trajectories accounted for 73.2% of the sample with the remaining 26.8% being summer trajectories. An expected side effect of this is that the seasonal classifiers show consistently lower errors due to the reduced variation in the samples they are trained with.

Figure 9 shows the 9-SOM trained on the summer data set (SUM-9), in this classifier, there is one Atlantic pathway (Class 1) and two visually smaller Atlantic pathways (Classes 4 and 7). The other classes making up this classifier

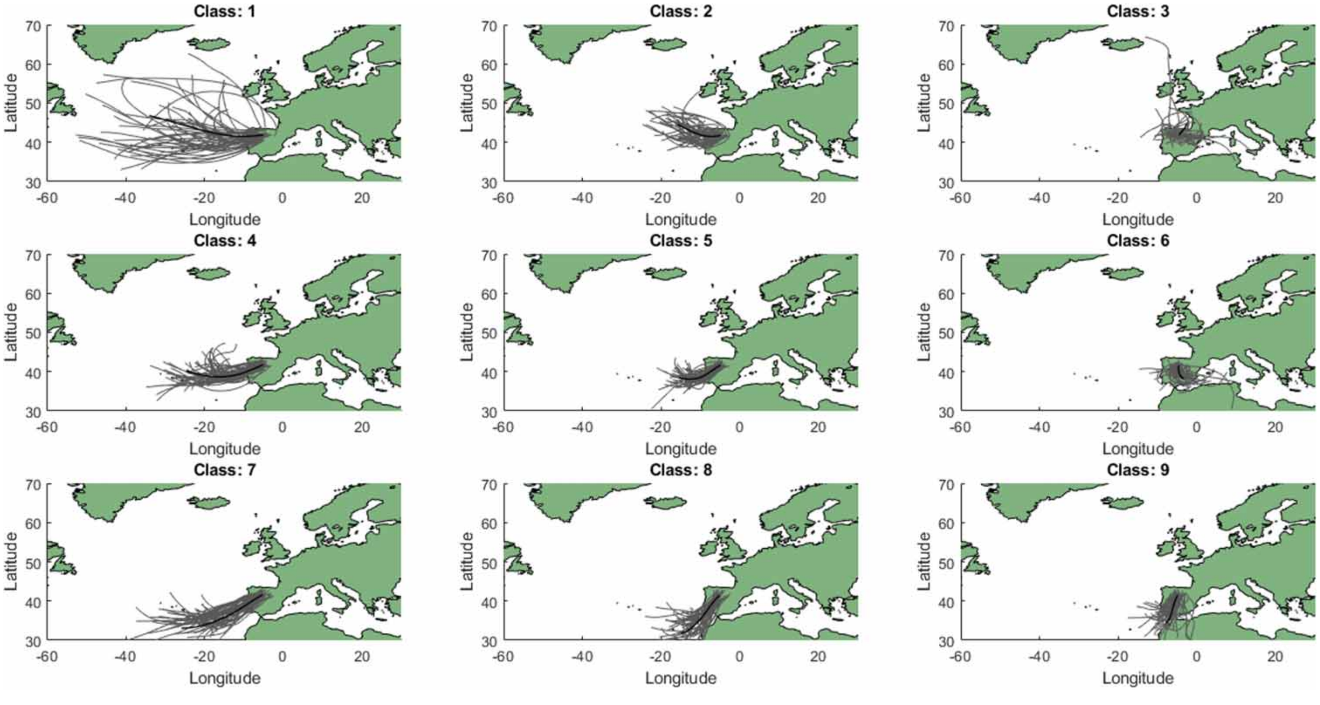

Figure 7 | A sample of trajectories for each cluster in the 9-SOM Primary classifier (PR-9). The classifications plot shows 50 randomly selected trajectories for each class (coloured grey) with a mean trajectory across all relevant trajectories coloured in black. The proportion of trajectories within each class is given in Table 3.

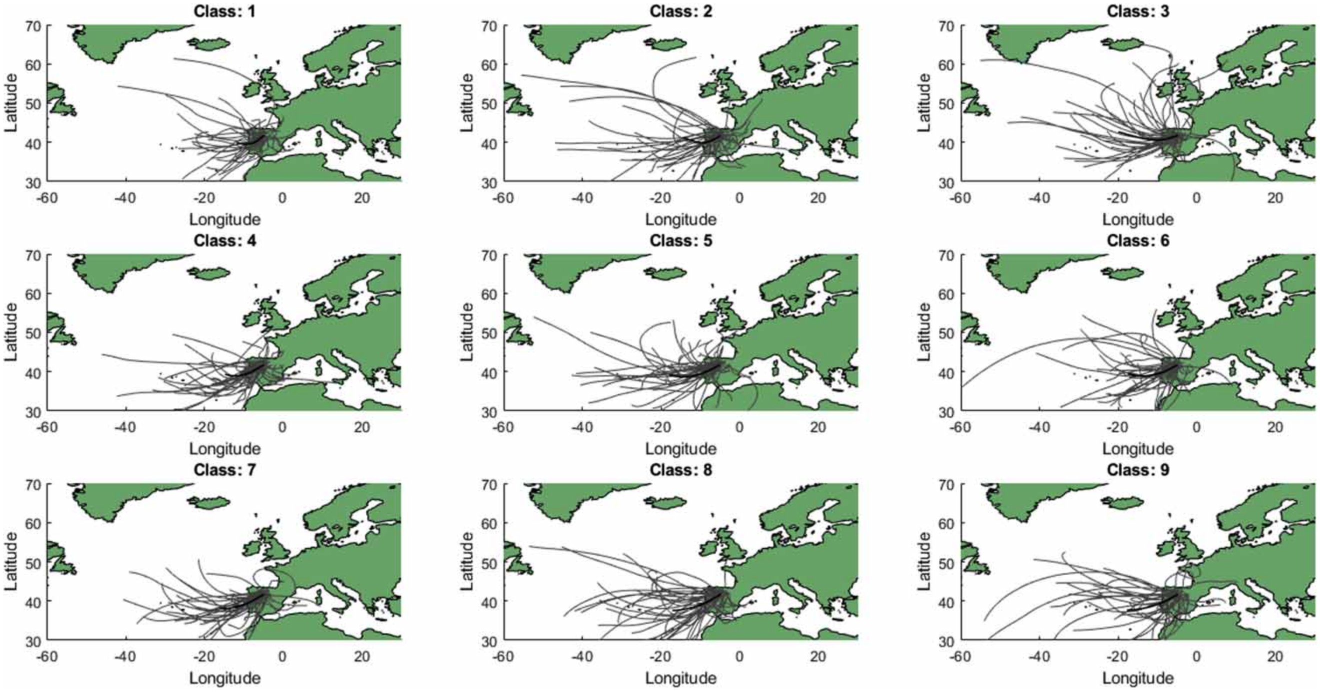

Figure 8 | A sample of trajectories for each cluster in the 3D 9-SOM classified (3D-9). The classifications plot shows 50 randomly selected trajectories for each class (coloured grey) with a mean trajectory across all relevant trajectories coloured in black. The proportion of trajectories within each class is given in Table 4.

Table 3 | Proportion of trajectories which are classified under each of the nine classes within PR-9

	Class								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
%	18.6	5.7	19.4	5.7	2.8	16.2	13.8	7.6	10.2

Table 4 | Proportion of trajectories which are classified under each of the nine classes within 3D-9

	Class								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
%	35.4	4.9	20.9	7.6	3.0	4.8	6.7	5.7	10.8

all visually appear as more refined continental storms, as discussed regarding PR-9 above. However, this classifier also presents three dominant pathways for the summer storms as given by Table 5, the mid-Atlantic (Class 1), northern continental (Class 3) and south-eastern Mediterranean tracks (Class 9). These pathways account for 16.8%, 17.5% and 20.7% (55% of the total pathways) of the summer tracks as shown in Table 5, reinforcing the conclusion from PR-9 that the dominant summer pathways are continental (or recirculatory) and east Atlantic originating tracks. These results also compare well with those found by Jorba *et al.* (2004) who as stated earlier identified recirculatory tracks as the dominant summer pathways; by separating these summer tracks we have further reinforced these results and shown that Atlantic tracks still play a key role in summer extremes.

The winter 9-SOM classifier (WIN-9) is shown in Figure 10 and appears visually similar to PR-9 which can be attributed to the winter storms comprising 73.2% of the sample. Dominant pathways in this classifier consist of deep-Atlantic (Class 1), continental (Class 3) and southern Atlantic (Class 7) paths; similarly, to SUM-9 these can be identified in earlier results such as Classes 1 and 7 in PR-9 and Class 1 in PR-4. These classes consist of 18.4%, 18.7% and 13.7% (50.8% of the total pathways) of the sample, which are significantly higher than Classes 2, 4 and 5 which only hold 6.3%, 6.8% and 3.6%. Further to this, 38.9% of winter events were classified (Classes 1, 4 and 7 in WIN-9) as deep-Atlantic compared to only 16.8% of summer events (Class 1 in SUM-9). This

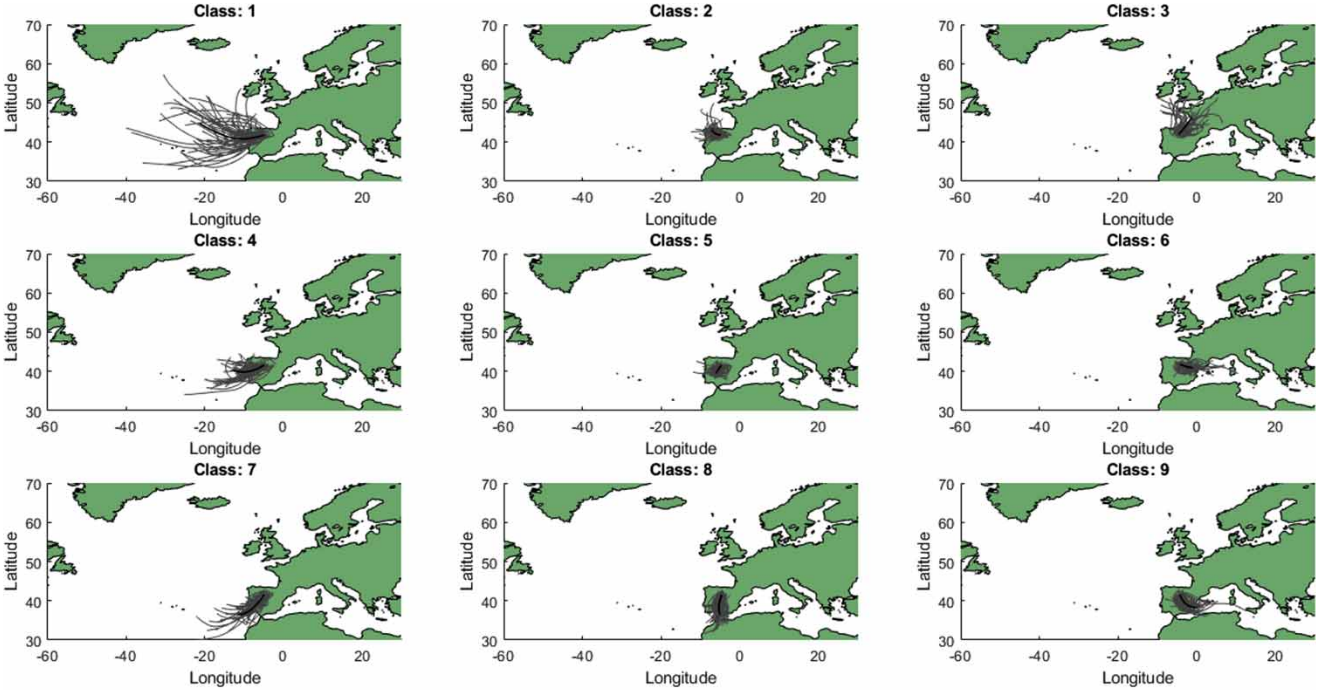

Figure 9 | A sample of trajectories for each class in the summer 9-SOM classified (SUM-9). The classifications plot shows 50 randomly selected trajectories for each class (coloured grey) with a mean trajectory across all relevant trajectories coloured in black. The proportion of trajectories within each class is given in Table 5.

Table 5 | Proportion of trajectories which are classified under each of the nine classes within SUM-9

	Class								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
%	16.8	7.0	17.5	4.8	5.3	10.7	9.2	8.0	20.7

shows deep-Atlantic storm trajectories have a seasonal dependence and generate at least twice as many AMAX events during the winter than during the summer.

Distributions of rainfall magnitude

A final analysis of the classifiers concerns the distributions of each cluster. Figure 11 shows the empirical distributions for each cluster in both PR-4 (a) and PR-9 (b). In both classifiers, the classes which are most likely to result in above average magnitude extreme events are from the southern Atlantic. For example, Class 3 from PR-4 and Class 7 from PR-9 have 21.7% and 23.4%, respectively, relatively of their tracks above this threshold. Both PR-4 and PR-9 also show similar results for the classes which are least likely to produce events above the threshold, with Class 2 (northern Atlantic) from both the PR-4 and PR-9 only having 18.7% and 18.6%, respectively. As these trajectories are similar this further reinforces the case that trajectories from the north Atlantic are less likely to cause the highest/lowest magnitude AMAX values but trajectories from the south Atlantic are more likely to cause these same events. Further to this, the proportions given in Tables 2 and 3 indicate that these north and south Atlantic trajectories are also the most uncommon.

Figure 12 shows the rainfall distributions for the classifications of both the summer (a) and winter (b) classifiers. The summer distributions show a separation into two groups most prominent at a normalised magnitude of 0.35. This separation occurs between the Atlantic tracks (Classes 1, 4 and 7) and the continental tracks (Classes 2, 3, 5, 6 and 8); Class 9 is different in that it does not join a group and instead holds a middle ground between the two. Taking a magnitude of 0.5 as a threshold, there is a maximum difference of 7.3% in the number of tracks with a magnitude which exceeds this threshold between the two groups (Classes 1, 2 and 6), the closest difference is 4.5% between

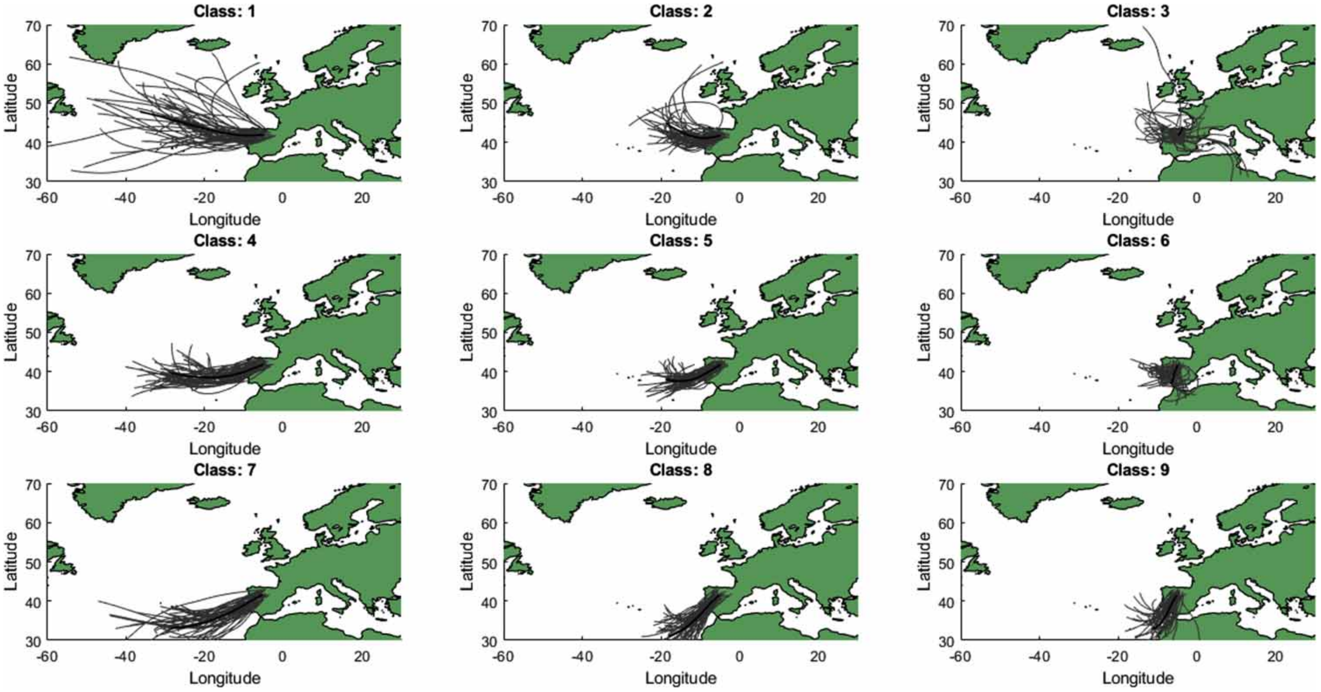

Figure 10 | A sample of trajectories for each cluster in the winter 9-SOM classifier (WIN-9). The classifications plot shows 50 randomly selected trajectories for each class (coloured grey) with a mean trajectory across all relevant trajectories coloured in black. The proportion of trajectories within each class is given in Table 6.

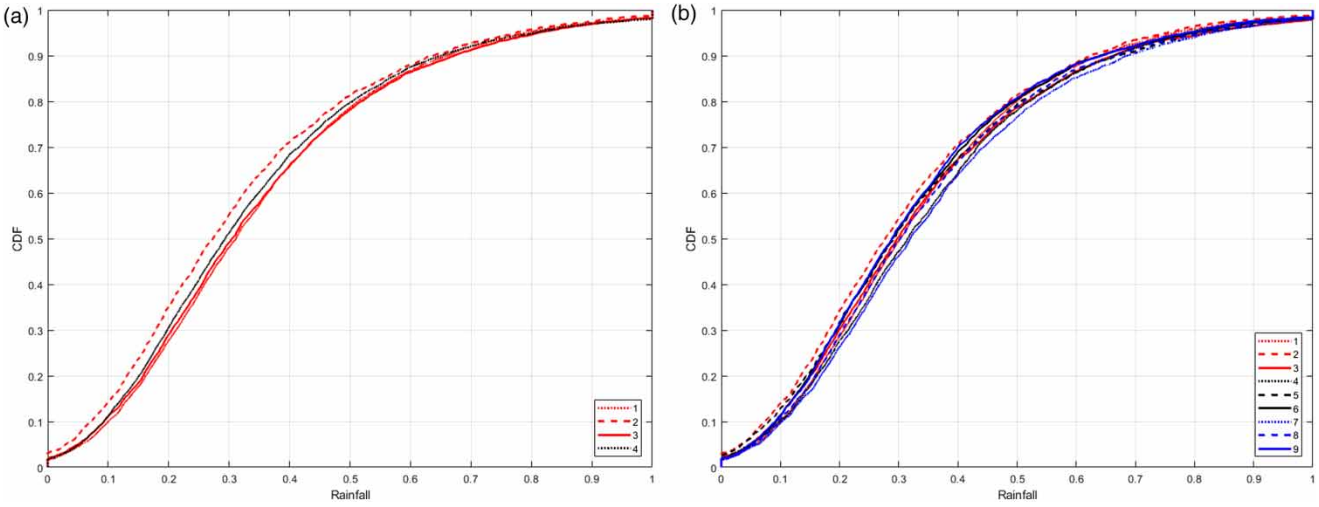

Figure 11 | Rainfall distributions for each cluster in PR-4 (a) and PR-9 (b). In each case, the line's number corresponds to the class of storm tracks represented.

Classes 7 and 8. This separation indicates that, during the summer months, it is continental tracks that are more likely to cause higher magnitude events.

In contrast, the winter distributions appear more varied with no dominant group separation as in the summer classifier. Taking the same experiment as above, using a magnitude threshold of 0.5, the lower Atlantic classes (Classes 7 and 5) have 23.8% and 24.0% of their samples exceeding this threshold. This indicates the lower Atlantic storms have a higher likelihood of causing above average magnitude AMAX events in the winter, and the continental storms have a

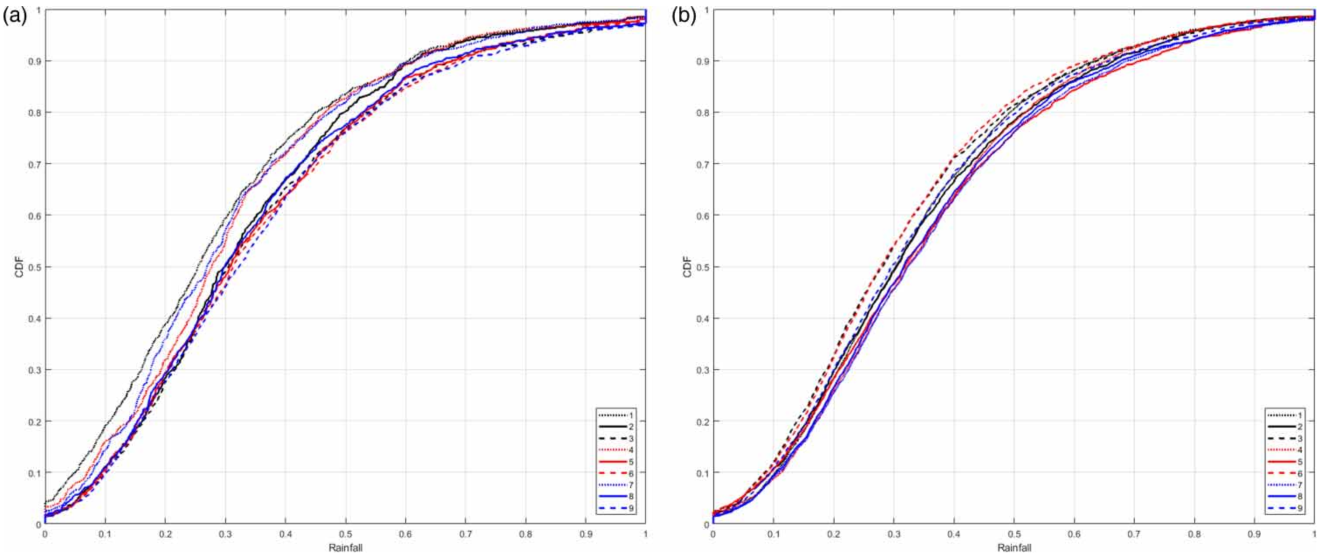

Figure 12 | Rainfall distributions for SUM-9 (a) and WIN-9 (b) classifiers. In each case, the line's number corresponds to the class of storm tracks represented.

Table 6 | Proportion of trajectories which are classified under each of the nine classifications within the 9-SOM winter classifier

	Class								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
%	18.4	6.3	18.7	6.8	3.6	12.5	13.7	8.2	11.8

higher likelihood of causing these same events in the summer.

CONCLUSIONS

This paper demonstrates the use of SOMs for improving the current understanding of rainfall frequency analysis through the classification of the moisture pathways leading to an AMAX rainfall event. Five classification models were generated using different subsets of the trajectory data in northern Spain, each provided insight into the applicability of the SOMs and the causes of extreme rainfall in this region. Here, we define extreme rainfall events as the AMAX of one-day precipitation measurements from a series of gauging stations.

1. Clustering on latitude/longitude (2D) trajectories for all rainfall events with four clusters (PR-4) and nine clusters (PR-9).
   a. Two prominent classes identified in PR-4 represent 81.7% of the sample trajectories; Class 1 from the mid-/western Atlantic Ocean (35.2%) and Class 4 containing recirculation patterns (46.5%).
   b. The prominent classes from PR-4 are verified by PR-9 in which the Atlantic pathways (Classes 1, 4 and 7) account for 38.0% of the sample trajectories.
   c. The southern Atlantic and mid-/western Atlantic pathways classes (Classes 1 and 3: PR-4) contained events which had 4.0% more high magnitude extreme events than the shorter Atlantic Class 2 as illustrated in Figure 11. A similar pattern is present in PR-9.
2. Clustering on altitude, latitude and longitude (3D) trajectories for all rainfall events.
   a. This model (3D-9) showed the highest numerical errors and produced clusters with little visual difference; this is due to the larger number of input variables required. This is often referred to as the curse of dimensionality.
3. Clustering on longitude and latitude (2D) trajectories for summer events (SUM-9).
   a. Three prominent pathways are responsible for 55% of summer extremes, these originate from: the mid-Atlantic (Class 1), northern Europe (Class 3) and the south-eastern Mediterranean (Class 9).

   b. The three Atlantic clusters (Classes 1, 4 and 7) contained at least 4.5% more high magnitude events than the other clusters; in some cases, this raised to 7.3%.

4. Clustering on longitude and latitude (2D) trajectories for winter events (WIN-9).

   a. The three dominant pathways in this model are similar to the primary models (PR-9 and PR-4) but only represent 50.8% of the trajectories: western Atlantic (Class 1: 18.4%), continental (Class 3: 18.7%) and southern Atlantic (Class 7: 13.7%).

   b. Southern Atlantic classes (Classes 5 and 7) are the most likely pathways to produce above average magnitude extremes at 24.0% and 24%, respectively.

These results show that clustering can provide improved insight into components of rainfall distributions. The SOM approach has proved capable of spatial clustering of trajectory patterns in 2D; however, we have also shown why care must be taken when considering 3D trajectories to minimise both numerical errors and visual difference. Finally, the results revealed differences in the origins of winter and summer extreme rainfall in the Douro region of northern Spain as detailed above. This work has opened up new questions in the use of alternative input variables to the clustering algorithm; for example, instead of normalising station data, a station type could be added as an input variable. Future studies could also investigate the development of a new clustering error metric accounting for both numerical errors and the number of clusters to limit the scope of training procedures.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model and the Spanish 'Agencia Estatal de Meteorología' (AEMET) for providing the precipitation data used. The first author acknowledges funding as part of the Water Informatics Science and Engineering Centre for Doctoral Training (WISE CDT) under a grant from the Engineering and Physical Sciences Research Council (EPSRC), grant no. EP/L016214/1.

REFERENCES

Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A. P., Merz, B., Arheimer, B. & Živković, N. 2017 Changing climate shifts timing of European floods. *Science* **357** (6351), 588–590.

Draxler, R. 1999 *HYSPLIT4 User's Guide*. NOAA Tech. Memo. ERL ARL-230, (June). NOAA Air Resources Laboratory, Silver Spring, MD.

Draxler, R. R. & Hess, G. D. 1997 *Description of the HYSPLIT4 Modeling System*. Tech Report, Silverspring, Maryland.

Draxler, R. R. & Hess, G. D. 1998 An overview of the HYSPLIT_4 modelling system for trajectories, dispersion, and deposition. *Australian Meteorological Magazine* **47**, 295–308.

Fahimi, F., Yaseen, Z. M. & El-shafie, A. 2016 Application of soft computing-based hybrid models in hydrological variables modeling: a comprehensive review. *Theoretical and Applied Climatology* **128** (3–4), 875–903.

Gimeno, L., Drumond, A., Nieto, R., Trigo, R. M. & Stohl, A. 2010a On the origin of continental precipitation. *Geophysical Research Letters* **37**, L13804. doi:10.1029/2010GL043712.

Gimeno, L., Nieto, R., Trigo, R. M., Vicente-Serrano, S. M. & Lopez-Moreno, J. I. 2010b Where does the Iberian peninsula moisture come from? An answer based on a Lagrangian approach. *Journal of Hydrometeorology* **11**, 421–436. doi:10.1175/2009JHM1182.1.

Hirschboeck, K. K. 1987 Hydroclimatically-defined mixed distributions in partial duration flood series. In: *Hydrologic Frequency Modeling* (V. P. Singh ed.). Springer, Dordrecht, pp. 199–212.

Jorba, O., Pérez, C., Rocadenbosch, F. & Baldasano, J. 2004 Cluster analysis of 4-day back trajectories arriving in the Barcelona area, Spain, from 1997 to 2002. *Journal of Applied Meteorology* **43** (6), 887–901.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. & Joseph, D. 1996 The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**, 437–472.

Kalteh, A. M., Hjorth, P. & Berndtsson, R. 2008 Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environmental Modelling & Software* **23** (7), 835–845.

Kiviluoto, K. 1996 Topology preservation in self-organizing maps. *Proceedings of International Conference on Neural Networks (ICNN'96)*. doi:10.1109/icnn.1996.548907.

Kjeldsen, T. R., Ahn, H., Prosdocimi, I. & Heo, J. H. 2018 Mixture Gumbel models for extreme series including infrequent phenomena. *Hydrological Sciences Journal* **63** (13–14), 1927–1940.

Kohonen, T. & Honkela, T. 2007 Kohonen network. *Scholarpedia Journal* **2** (1), 1568.

Lee, J.-G., Han, J., Li, X. & Gonzalez, H. 2008 TraClass. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* **1** (1), 1081–1094.

Marimont, R. B. & Shapiro, M. B. 1979 Nearest neighbour searches and the curse of dimensionality. *IMA Journal of Applied Mathematics* **24** (1), 59–70.

Mediero, L., Santillan, D., Garrote, L. & Granados, A. 2014 Detection and attribution of trends in magnitude, frequency and timing of floods in Spain. *Journal of Hydrology* **517**, 1072–1088. doi:10.1016/j.jhydrol.2014.06.040.

Merz, R. & Blöschl, G. 2003 A process typology of regional floods. *Water Resources Research* **39**, 12. doi:10.1029/2002wr001952.

NOAA 2003 *NCEP/NCAR Global Reanalysis Data Archive Information*. Available from: https://www.ready.noaa.gov/gbl_reanalysis.php (accessed 11 September 2019).

Owens, J. & Hunter, A. 2000 Application of the self-organising map to trajectory classification. *Proceedings Third IEEE International Workshop on Visual Surveillance*. doi:10.1109/vs.2000.856860.

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.

Scoccimarro, E., Gualdi, S. & Krichak, S. 2018 Extreme precipitation events over north-western Europe: getting water from the tropics. *Annals of Geophysics* **61** (4). https://www.annalsofgeophysics.eu/index.php/annals/article/view/7772

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.

Stohl, A. & James, P. 2004 A Lagrangian analysis of the atmospheric branch of the global water cycle. Part I: method description, validation, and demonstration for the August 2002 flooding in central Europe. *Journal of Hydrometeorology* **5** (4), 656–678.

Tan, X., Gan, T. Y. & Chen, Y. D. 2017 Moisture sources and pathways associated with the spatial variability of seasonal extreme precipitation over Canada. *Climate Dynamics* **50** (1–2), 629–640.

Utsumi, N., Kim, H., Kanae, S. & Oki, T. 2016 Relative contributions of weather systems to mean and extreme global precipitation. *Journal of Geophysical Research: Atmospheres* **122**, 152–167. doi:10.1002/2016JD025222.

Villarini, G. & Smith, J. A. 2010 Flood peak distributions for the eastern United States. *Water Resources Research* **46**, 6. doi:10.1029/2009WR008395.

Wallace, J. M., Wallace, J. M. & Hobbs, P. V. 1977 *Atmospheric Science: An Introductory Survey*, Elsevier Academic Press, Amsterdam, The Netherlands.

Waylen, P. & Woo, M. K. 1982 Prediction of annual floods generated by mixed processes. *Water Resources Research* **18** (4), 1283–1286.

Zhuang, X., Huang, Y., Palaniappan, K. & Zhao, Y. 1996 Gaussian mixture density modelling, decomposition, and applications. *IEEE Transactions on Image Processing* **5** (9), 1293–1302.

First received 22 March 2019; accepted in revised form 7 October 2019. Available online 23 October 2019