

A Generalized Density-Based Algorithm for the Spatiotemporal Tracking of Drought Events

C. CAMMALLERI^a AND A. TORETI^a

^a *Joint Research Centre, European Commission, Ispra, Italy*

(Manuscript received 6 June 2022, in final form 10 November 2022)

ABSTRACT: Drought events evolve simultaneously in space and time; hence, a proper characterization of an event requires the tracking of its full spatiotemporal evolution. Here we present a generalized algorithm for the tracking of drought events based on a three-dimensional application of the DBSCAN (density-based spatial clustering of applications with noise) clustering approach. The need for a generalized and flexible algorithm is dictated by the absence of a unanimous consensus on the definition of a drought event, which often depends on the target of the study. The proposed methodology introduces a set of six parameters that control both the spatial and the temporal connectivity between cells under drought conditions, also accounting for the local intensity of the drought itself. The capability of the algorithm to adapt to different drought definitions is tested successfully over a study case in Australia in the period 2017–20 using a set of standardized precipitation index (SPI) data derived from the ERA5 precipitation reanalysis. Insights on the possible range of variability of the model parameters, as well as on their effects on the delineation of drought events, are provided for the case of meteorological droughts in order to incentivize further applications of the methodology.

KEYWORDS: Drought; Precipitation; Clustering

1. Introduction

Drought is a climate extreme that can occur in all climate zones of the world and that has serious impacts on society and the environment (Mishra and Singh 2010). In recent years, impacts of droughts have been exacerbated by human activities and water exploitation, resulting in renovated interest in drought risk management, reduction, and mitigation (UNDRR 2021). Our capability to analyze and characterize the spatiotemporal dynamics of drought events from indicators, derived from either observations or model predictions, is instrumental in better understanding the dynamics and the consequent cascading repercussions of droughts on socioeconomic and environmental sectors.


Even if drought is widely recognized as a spatiotemporal phenomenon (Tallaksen and Van Lanen 2004), it is rarely properly treated as such, with most previous studies reducing the analysis to investigating the temporal variability (time series) for given predefined areas or to studying the spatial patterns for a fixed period (time window) (i.e., González-Hidalgo et al. 2018; Spinoni et al. 2015; Zhao et al. 2017).

One of the earliest attempts at developing an algorithm for the spatiotemporal tracking of drought events can be found in Andreadis et al. (2005). This pioneering work introduced the definition of contiguous drought area (CDA) to aggregate

spatial cells into drought clusters, as well as the possibility for clusters to merge and split at subsequent time steps. Sheffield et al. (2009), Herrera-Estrada et al. (2017), and Diaz et al. (2020) employed similar tracking procedures, keeping at the core the same CDA analysis but also exploring different strategies for the temporal tracking component, including evaluations on the effects of minimum cluster size and of the distance between clusters.

By extending the CDA analysis to the full space–time domain, Lloyd-Hughes (2012) suggested to directly derive coherent spatiotemporal structures, merging de facto the previous two steps of clustering and tracking into a single three-dimensional clustering. Vernieuwe et al. (2020) and Diaz et al. (2023) further elaborated on this direction by investigating mathematical morphology and parameter optimization approaches. In spite of the understandable advantages of a single three-dimensional procedure, Lloyd-Hughes (2012) highlighted how this approach implicitly equates the dimensions along the space (i.e., longitude and latitude) and the time scales, and that scaling factors may be required to account for inequality. In addition, since 3D clustering requires a fixed 3D dataset (longitude, latitude, and time) to provide a stable outcome, some of the advantages of 3D clustering over a 2D clustering combined with a temporal tracking can be lost in a near-real-time application. Indeed, the outcome of the 3D cluster may change in time when new data are added in a continuously evolving dataset, e.g., used for monitoring and early warning purposes.

A major limitation for the application of these procedures is the well-known lack of consensus on a unanimous definition of drought event. It is not uncommon that different experts may characterize the same event differently based on their knowledge, or even that the same expert may use different definitions depending on the goal of the study. Multiple drought indices exist to accommodate for the large variety of

 Denotes content that is immediately available upon publication as open access.

Cammalleri's current affiliation: Politecnico di Milano, Milan, Italy.

Corresponding author: C. Cammalleri, carmelo.cammalleri@polimi.it

DOI: 10.1175/JHM-D-22-0115.1

© 2023 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy \(www.ametsoc.org/PUBSReuseLicenses\)](https://www.ametsoc.org/PUBSReuseLicenses).

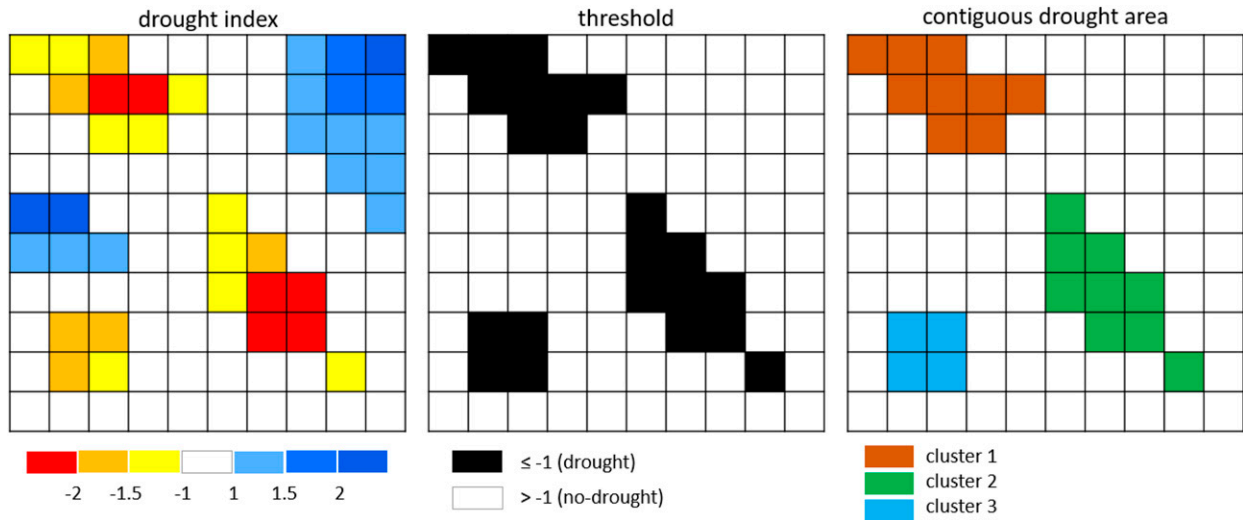


FIG. 1. Schematic representation of the contiguous drought area (CDA) procedure.

potential applications and spatial scales (WMO and GWP 2016), and a similar flexibility may be needed in drought tracking algorithms. As an example, while short-term standardized precipitation indices (SPIs; i.e., SPI-1 and SPI-3) are more suitable for detecting drought events that are relevant for meteorological and agricultural applications, long-term accumulation periods (i.e., SPI-12 or longer) address the needs of climatological and hydrological studies. Analogously, climatological studies may be more inclined to track long-lasting droughts related to multiyear abnormally dry conditions, whereas a splitting of these major droughts into subevents may be of interest for monitoring applications.

Starting from the approaches already available in the literature, the goal of this study is to introduce a generalized density-based method that allows for an increased flexibility in the characterization of spatiotemporal drought events. The proposed algorithm is detailed by highlighting its main assumptions with respect to the existing methodologies, and then by describing the expected behavior in cluster duration and extension based on different settings of its parameters. As a first case study, a drought period in Australia during the late 2010s is investigated, with the goal to confirm the capability of the approach to flexibly adapt to different characterizations of the same drought period.

2. Methodology

Starting from the work of Andreadis et al. (2005), most drought tracking algorithms rely on the detection of contiguous drought areas (CDA) for a series of time steps, which are then converted into drought events in a second tracking phase. Here we introduce a generalization of the CDA procedure for the 2D clustering, and then we highlight how to extend this generalized approach to the 3D case.

a. Generalized 2D clustering

The procedure adopted to identify the CDAs starting from a drought indicator map is schematized in Fig. 1 as a two-step

process. First, the drought indicator is converted into a Boolean map of drought/no-drought cells based on a given threshold. In Fig. 1, the commonly used threshold value of -1 is adopted to discriminate drought conditions in a hypothetical SPI map. Then, contiguous cells (using an 8-neighborhood rule) are grouped into CDAs, also known as clusters. It is worth noting how some of these clusters may be considered too small to be tracked as drought events, hence they are successively filtered through a minimum area value, as detailed in section 2b.

The above-described algorithm is based on a rather simplistic aggregation approach that does not account for much flexibility in the cluster mapping (given that the only tunable parameter is the threshold value). In the framework of machine learning, several types of clustering techniques have been introduced (see, e.g., Madhulatha 2012); among them, there is the density-based class, such as the well-known density-based spatial clustering of applications with noise (DBSCAN; Ester et al. 1996) method. DBSCAN has been deployed in a large variety of applications, including studies on meteorological extremes (Tilloy et al. 2022; Wang and Yan 2021).

The DBSCAN method exploits the distance between points to turn dense areas of data points into clusters. It groups together points according to their proximity and data density. Data that are packed in high-density regions are defined as clusters, whereas points in low-density areas are marked as outliers. The algorithm distinguishes between three categories:

- core points, having at least a minimum number of points in their surroundings;
- reachable points, lying within a minimum distance from a core point; and
- outliers, being not reachable from any other point.

In its general formulation, DBSCAN requires only two main parameters to generate the clusters: 1) a searching

radius for the detection of neighborhood points (eps), and 2) a minimum number of points (minPts) in the neighbor to classify a point as a part of the cluster core. Among the advantages of DBSCAN there is the fact that a preset number of clusters is not needed, and that arbitrarily sized and shaped clusters can be detected. Known limitations of the algorithm include the complexity to define the searching radius if the data scale is not well understood, and the difficulty to cluster data with large differences in densities (Kriegel et al. 2011).

In this study, a weighted DBSCAN implementation is adopted (Hahsler et al. 2019); the weighting factor (between 0 and 1) is attributed to each cell to either increase or decrease the weight of a point in forming a cluster. Weighting was originally introduced to account for the presence of multiple samples at the same location, in order to remove duplicates and speed up the computational time (by including a single point in the analysis but giving it a higher weight based on the number of replicates). However, it can clearly be also used to give different weights to different values of the drought indicator rather than adopting a simple Boolean classification (i.e., all the points weight the same).

In the special case of data distributed on a regular grid, it is conceptually easier to replace the search radius parameter eps with the length of a searching window L . Similarly, on a regular grid, the values assumed by minPts will be always bounded between a minimum of 2 (one additional point in the searching area beyond the one under analysis) and the total number of points in the searching window (e.g., 9 in the case of a 3×3 window). Hence, the parameter minPts can be replaced by a parameter p (between 0 and 1) representing the fraction between the minimum and maximum values for a given window.

Regarding the weighting factor w , following the approach adopted by Cammalleri et al. (2016) for the quantification of the dryness of a standardized drought index, a simple logistic function is here adopted, such as

$$w = \frac{1}{1 + \left(\frac{D}{k}\right)^e}, \quad (1)$$

where D is the drought index, k is the D value at which the logistic function is centered, and e is the steepness of the growth curve. In the case of a standardized quantity (i.e., SPI), this function is applied to only one side of the range of D , while $w = 0$ is used for the other half of the range.

In this generalized formulation, the clustering procedure becomes a function of four parameters: L and p related to the classical DBSCAN, and k and e linked to the weighting function. This formulation represents a clear step up in flexibility compared to the single parameter (threshold) of the classical CDA formulation, as detailed in section 4a. The classical CDA approach proposed by Andreadis et al. (2005) and described in Fig. 1 can be reproduced as a special case of this general framework by setting $L = 3$, $p = 0$ (corresponding to $\text{eps} = \sqrt{2}$ and $\text{minPts} = 2$), $k = -1$, and $e = 1000$ (the latter being a large enough value to approximate the Boolean function with an S-shaped curve).

b. Extension to 3D clustering

As briefly detailed in the introduction, several different approaches have been adopted in drought studies to track the temporal evolution of 2D spatial clusters retrieved at different time steps, e.g., linking clusters via simple overlaps (Herrera-Estrada et al. 2017), evaluating the distance between centroids (Diaz et al. 2020), or applying proper 3D clustering (Lloyd-Hughes 2012).

As introduced in section 2a, one common step in most of the clustering procedures is the use of a minimum cluster area (A) to filter out the small aggregations of grid cells that could represent tenuous spatial connections between otherwise unrelated events (Sheffield et al. 2009). This filtering can be applied to the proposed DBSCAN 2D clustering as a simple postprocessing step, after the clusters of all sizes are produced and stored for each time step (e.g., the cluster 3 in Fig. 1 would be filtered out if $A = 5$ is used).

The 2D clustering framework introduced in section 2a (applied to the latitude–longitude data for a given time step) can be immediately extended in 3D by simply stacking the data acquired for all the available time steps, thus obtaining a so-called data cube (a full dataset comprising all the available data in a latitude–longitude–time 3D spatial domain). This extension assumes that the searching area for neighborhood cells is not a circle anymore but rather a sphere. However, as pointed out by Lloyd-Hughes (2012), such an approach implicitly assumes that the length scales of space (i.e., latitude and longitude) and time (i.e., time steps) are comparable, hence the same search window used in space (L) would be adopted also in time. As a practical example, this symmetric search window approach assumes that two points that are separated by a single unit in space (e.g., 1°) are considered “equidistant” to two points that are separated by a single unit in time (e.g., 1 month). To overcome this limitation, and enable an asymmetric spatiotemporal search window, a scaling factor for the time dimension can be introduced, thus defining (in analogy to the space scale) the length of the search window in time, R . The distinction between the size of the search window for space (L) and time (R) allows modulating the algorithm behavior for different combinations of spatial and temporal resolutions.

The full 3D formulation is then applied as synthetically schematized in Fig. 2. First, a preliminary application of the 2D weighted DBSCAN is performed independently for each time step in the period under analysis (dotted box in Fig. 2). This preliminary step is used to filter the clusters with size smaller than A at each given time step, and it produces a full data cube comprising only the unfiltered grid cells characterized by the four quantities: weight, latitude, longitude, and (scaled) time step. In the second phase, the proper 3D clustering is performed on the previously generated data cube by using the same parameters adopted in the 2D DBSCAN.

The proposed generalized algorithm has six parameters (the four outlined in section 2a plus A and R), adding further flexibility to the methodology in treating the connectivity between drought cells in the 3D space.

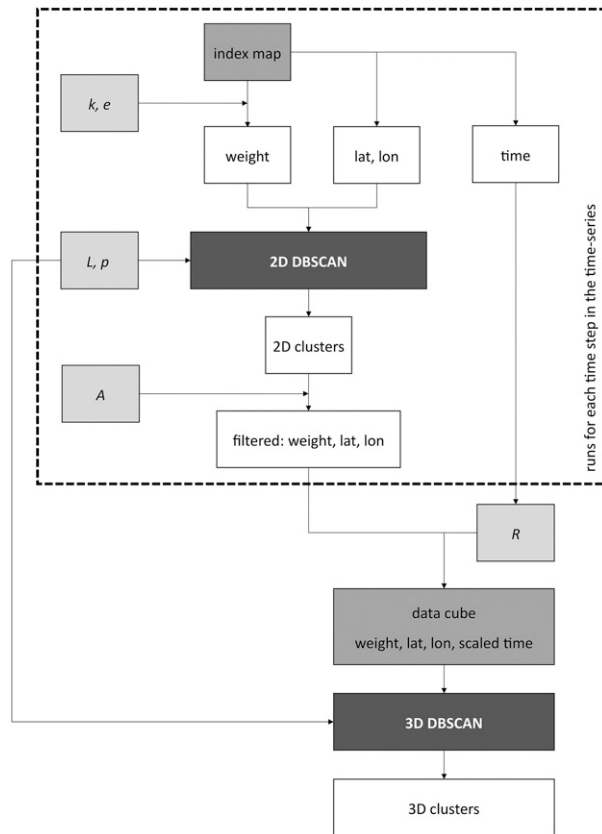


FIG. 2. Schema of the proposed 3D clustering procedure. The preprocessing step performs a 2D clustering for each time step (enclosed in the dotted box) to filter the data, followed by the proper 3D clustering on the full data cube.

3. Data

To test the proposed clustering approach, a global dataset of SPI-3 (i.e., 3 months accumulation) is here mainly used. This dataset is derived from the ERA5 precipitation reanalysis (Hersbach et al. 2020), as fully described in Cammalleri et al. (2022), and it covers the period 1981–2020 on a regular latitude–longitude grid having a spatial resolution of 0.25° with temporal updates of roughly 10 days (i.e., dekad, three updates per month on the 10th, 20th, and the last day of each month).

The SPI is estimated by fitting a gamma distribution on a set of multiyear precipitation data by using a reference period of 30 years (1981–2010), as suggested by the WMO guidelines (WMO 2017). The two parameters of the gamma distribution are derived separately for each dekadal time series using the generalized additive model for location, scale, and shape (GAMLSS) modeling framework (Stasinopoulos and Rigby 2007). A total of 36 gamma fittings (12 months \times 3 monthly updates) are produced, and the complete SPI dataset derived from these fittings consists of 1440 global maps.

A time series of SPI-6 (6 months accumulation) maps is also computed, using the same ERA5 precipitation dataset

but with monthly temporal updates, for a total of 480 global maps. This dataset is used in addition to SPI-3 to test how the clustering algorithm performs with the same parameterization but different input data.

4. Results and discussion

a. Parameterization of the clustering algorithm

Having introduced a set of six parameters, it is important to understand how each of them influences the outcome of the clustering analysis.

Regarding the classical DBSCAN parameters (viz., L and p), it is easy to understand how they are closely interlinked, so the effects of these two are jointly analyzed. While the window search size can assume a wide range of values, large L values quickly start to connect areas too far away. For moderate-resolution datasets, like ERA5, Vernieuwe et al. (2020) considered three window sizes, (3×3 , 5×5 , and 7×7) which already cover a rather substantial range of variability for the L parameter. In general (for a given p value), larger values of L correspond to bigger clusters and an increase in aggregation of drought cells into fewer clusters. The CDA represents the lower degree of aggregation, since $L = 3$ is the minimum admissible value.

On the contrary, an increase in p results in more fragmented clusters (for a fixed value of L), since less core cells will be identified due to the higher constraint on the number of pixels in the surrounding required to identify a core cell. Even if p can go up to 1 (i.e., all the cells in the surrounding need to be under drought conditions to flag a core cell), an already quite extreme constraint is to assume p equal to 0.5, i.e., a point is flagged as core when at least half of the surrounding cells are under drought conditions. For this parameter, the CDA represents the maximum possible aggregation (for $L = 3$), given that a single cell in the surrounding is enough to guarantee the clustering.

These suggested ranges for L (between 3 and 7) and p (between 0 and 0.5) already provide a wide array of options, with few examples shown in Fig. 3 for SPI-3 in a case over the United States. It is worth highlighting that these examples are obtained with the simple Boolean weighting method, just to provide a general overview of the expected qualitative behavior.

Figure 3 shows the clustering with four different parameterizations on a sample dataset. The map in Fig. 3a depicts the results for the CDA, as detailed at the end of section 2a, while an increase in the searching window ($L = 7$, Fig. 3b) returns larger clusters also comprising nonadjacent cells. The lower panels show results for an intermediate L value ($L = 5$) by setting either a low (0.15, Fig. 3c) or a high (0.4, Fig. 4d) p value. These results highlight how multiple outcomes can be obtained by tuning L and p ; although, with the simple Boolean weighting it is unusual to get the major clusters split into subclusters (which can be instead achieved by introducing alternative weighting functions).

The logistic function introduced in Eq. (1) hence aims at giving further flexibility also in the clustering strategy for large areas, allowing for reducing the effects of weak

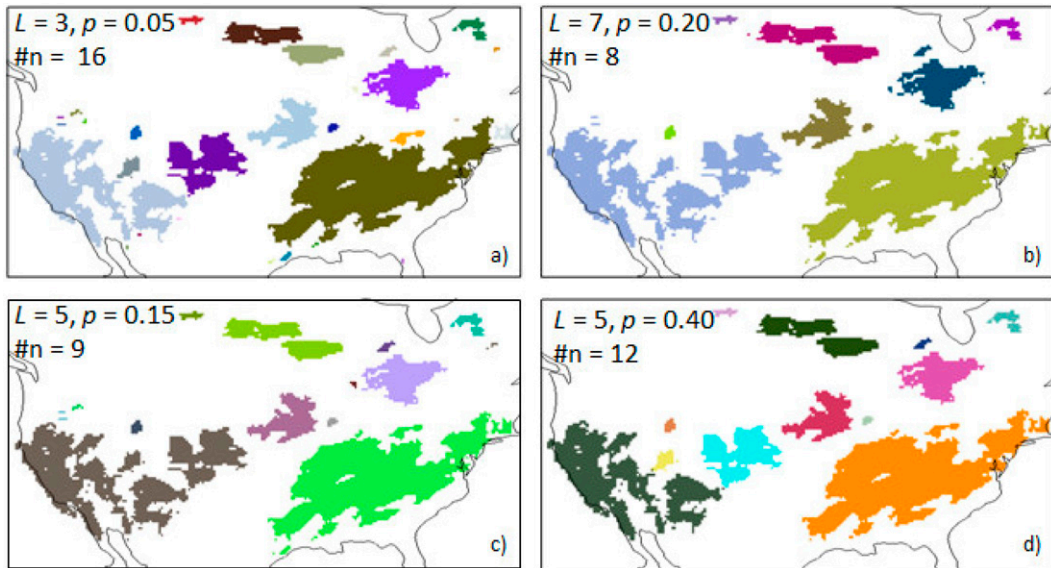


FIG. 3. Examples of 2D clustering based on different pairs of L and p parameters for a case study over the United States. A Boolean weighting is used in all four panels. The value of “#n” represents the number of clusters produced by the algorithm (only events larger than 10 cells are counted here).

connectivity due to small areas close to the threshold value. Following [Cammalleri et al. \(2016\)](#) and accounting for the frequency associated to different SPI values, k can be set reasonably between -1.2 and -1.8 , while e between 4 and 10. The maps in [Fig. 4](#) clearly reveal the effect of the logistic weighting factor for a case study over Europe. A single large cluster ([Fig. 4c](#)) detected by the classic CDA (i.e., Boolean weight) is split into two subclusters ([Fig. 4d](#)). The split occurs in correspondence of a weak link (having SPI values close to -1 over

Russia) between two major drought areas with SPI values well below -2 .

It is worth mentioning that the proposed weighting function (and the suggested range of variability) is suitable for standardized drought indices, i.e., centered on 0 and with a unitary standard deviation. Indicators with a different range (e.g., the Palmer drought severity index; [Alley 1984](#)) may need an ad hoc weighting function, based on the same principle to keep the 0–1 range.

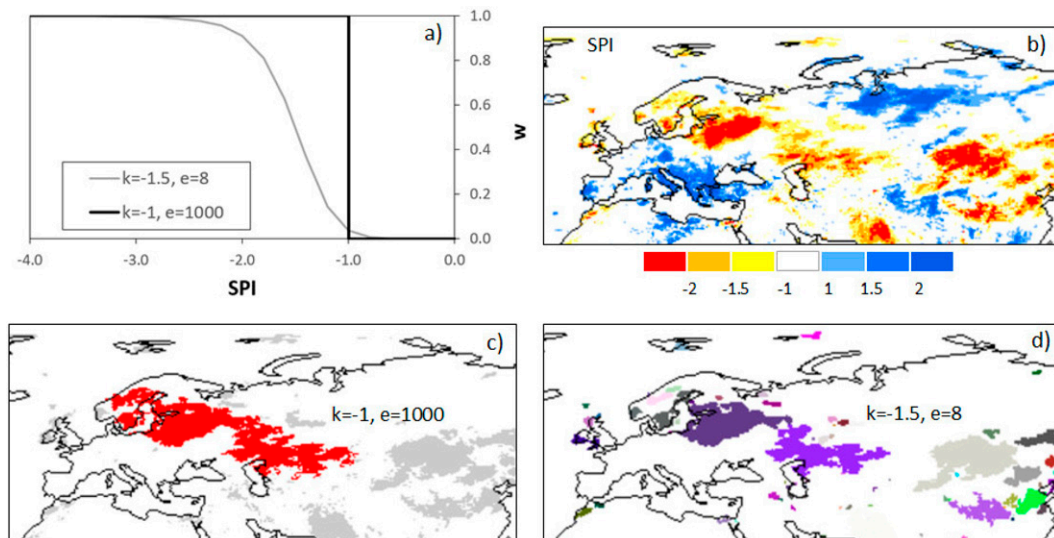


FIG. 4. Effect of the weighting function on the clustering for a case study over Europe. (a) The Boolean (black line) and logistic (gray line) weighting functions; (b) the SPI-3 input data to be clustered. (c),(d) The outcome of the clustering algorithm for the Boolean and logistic weighting, respectively. Both maps were obtained with $L = 3$ and $p = 0.05$.

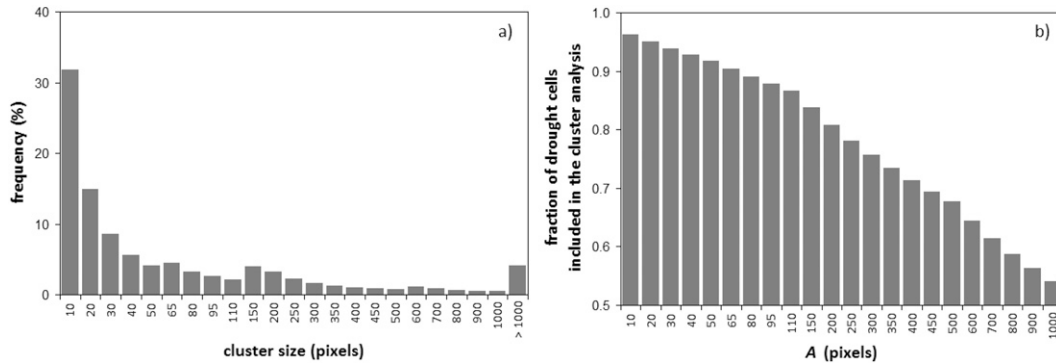


FIG. 5. Analysis of the optimal range of variability of the A parameter. (a) The frequency distribution of the number of clusters for different sizes based on a range of DBSCAN parameterizations. (b) The average ratio between the number of cells within clusters greater than A and the total number of cells considered under drought ($\text{SPI-3} < -1$).

Independently from the selected parameterization, DBSCAN produces clusters of various size (greater than minPts), so the A parameter plays the role of noise filter. Keeping all the small clusters in the analysis will result in larger and longer drought events, since these small areas act as connections between larger clusters. Thus, the filter aims at reducing the size of the final modeled drought events, as highlighted by Sheffield et al. (2009).

Various values have been adopted in the literature for such a parameter, e.g., Herrera-Estrada et al. (2017) used $200\,000\text{ km}^2$ to filter smaller clusters, Andreadis et al. (2005) used $25\,000\text{ km}^2$ over the United States, and Sheffield et al. (2009) used $500\,000\text{ km}^2$ for global applications (after unsuccessful initial experiments with both $25\,000\text{ km}^2$ and $100\,000\text{ km}^2$). Recently, Rakovec et al. (2022) used a threshold of $25\,000\text{ km}^2$ to filter small clusters, but in combination with a minimum temporal overlap between consecutive clusters of $150\,000\text{ km}^2$. Diaz et al. (2023) developed a strategy to detect a minimum value for A , based on the plateau of the cluster size curve. Over South America, their analysis returned a value of 49 pixels (corresponding to about $40\,000\text{ km}^2$) as a minimum cluster area.

This overview shows how lower A values seem to have been adopted in continental scale studies (from $25\,000$ to $40\,000\text{ km}^2$), where the limited extent of the study domain reduces the risk of getting events that persist (incorrectly) for multiple years spreading over multiple continents (see Sheffield et al. 2009). Large values of A can reduce these effects, but too large values may exclude from the cluster output a large fraction of the cells that are considered under drought conditions.

Another role of the A parameter is to control the onset and end of an event, given that an event starts when its size exceeds the selected A values, and it terminates when the cluster size reduces below this threshold. As a general rule of thumb, high values of A correspond to shorter events (late start and/or early end).

To identify an optimal range of variability for this parameter, an analysis of the frequency of clusters of different size was performed for our global SPI-3 dataset with six different L and p combinations (for both the Boolean weighting and a specific logistic weighting with $k = -1.5$ and $e = 6$), as

summarized by the results shown in Fig. 5a. A flexing point occurs between 50 and 80 pixels, a result roughly in line with the outcome of Diaz et al. (2023). On the other hand, Fig. 5b shows that for an A value in the order of 250 pixels (about $200\,000\text{ km}^2$) more than 20% of the cells considered under drought ($\text{SPI-3} < -1$) would be excluded by the clustering analysis. Following these two results, we consider $50\,000$ – $200\,000\text{ km}^2$ a reliable range of variability for global analyses (roughly between 65 and 250 pixels in our ERA5-based 0.25° dataset).

Finally, the R parameter plays exactly the same role for time as the L parameter does for space, hence it increases the distance where connections are detected along the temporal line. When R and L are set to be equal, the algorithm assumes

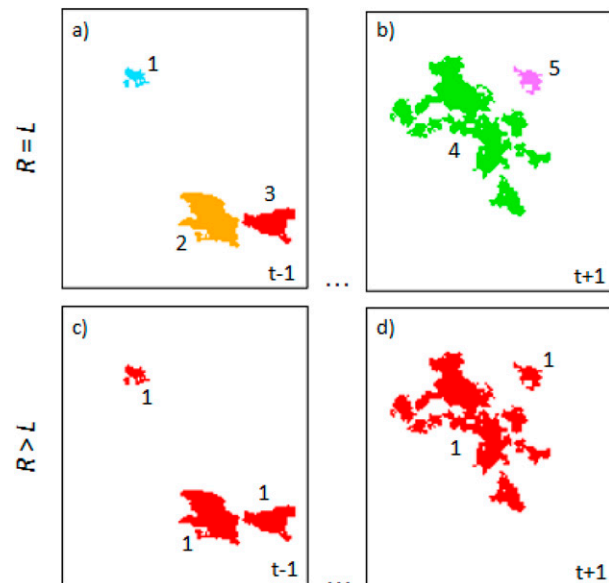


FIG. 6. Effects of missing data (empty map at time t) on temporal clustering: (a),(b) associated with a symmetric searching window ($R = L = 3$) and (c),(d) showing larger temporal searching window ($L = 3$ and $R = 5$). Each color represents a different event according to the specific clustering parameterization. The number next to each group of cells represents the ID of the event.

a symmetry in the searching window in the space–time domain. Consequently, this parameter can also be used to account for the inequality between the spatial and the temporal scales.

This asymmetry can be used, as an example, to account for gaps in the time series even without altering the continuity assumption in space. Figure 6 shows how an empty map at time t can disconnect the clusters at $t - 1$ (Fig. 6a) and at $t + 1$ (Fig. 6b) when $R = L = 3$ (i.e., only the closest time step is used to link clusters). In contrast, a connection is established when a larger temporal searching window is used ($R = 5$) even if the spatial searching window is kept the same (Figs. 6c,d).

Generally, in analogy to L , an increase in R results in bigger clusters (i.e., larger spatial extent and longer duration) for a given set of static parameters. The classical CDA application followed by an overlap represents in this case the minimum degree of aggregation, since only the closest time steps are considered. The use of two independent parameters in space in time allows for extending the search in one dimension without altering the assumption adopted in the other. This property can be used to either extend the duration of an event by connecting multiple consecutive clusters separated by small

TABLE 1. Summary of the suggested range of variability for the parameters of the algorithm. The values in parentheses correspond to the special case of the CDA with Boolean weighting. Parameters: L is the size of the searching window in space; p is the fraction of cells in the surrounding to detect a core cell; k is the midpoint of the logistic curve; e is the steepness of the logistic curve; A is the minimum size of the 2D clusters; and R is the size of the searching window in time.

| Parameter | L | p | k | e | A | R |
|-----------|-------|----------|-----------|----------|-----|-----|
| Min | 3 (3) | 0.05 (0) | -1.2 (-1) | 4 | 65 | 3 |
| Max | 7 | 0.5 | -1.8 (-2) | 8 (1000) | 250 | 7 |

temporal gaps (by increasing R) or by splitting long events into multiple subevents (by decreasing R). Although R acts on the time scale, it can also affect the spatial outcome, as exemplified by the decrease in the number of independent clusters obtained for a given time step in Fig. 6 (3 clusters in Fig. 6a for $R = 3$ and only 1 cluster in Fig. 6c for $R = 5$).

These findings on the algorithm behavior allow us to derive plausible ranges of variability, as summarized in Table 1. Those values were derived for global SPI-3 moderate-resolution

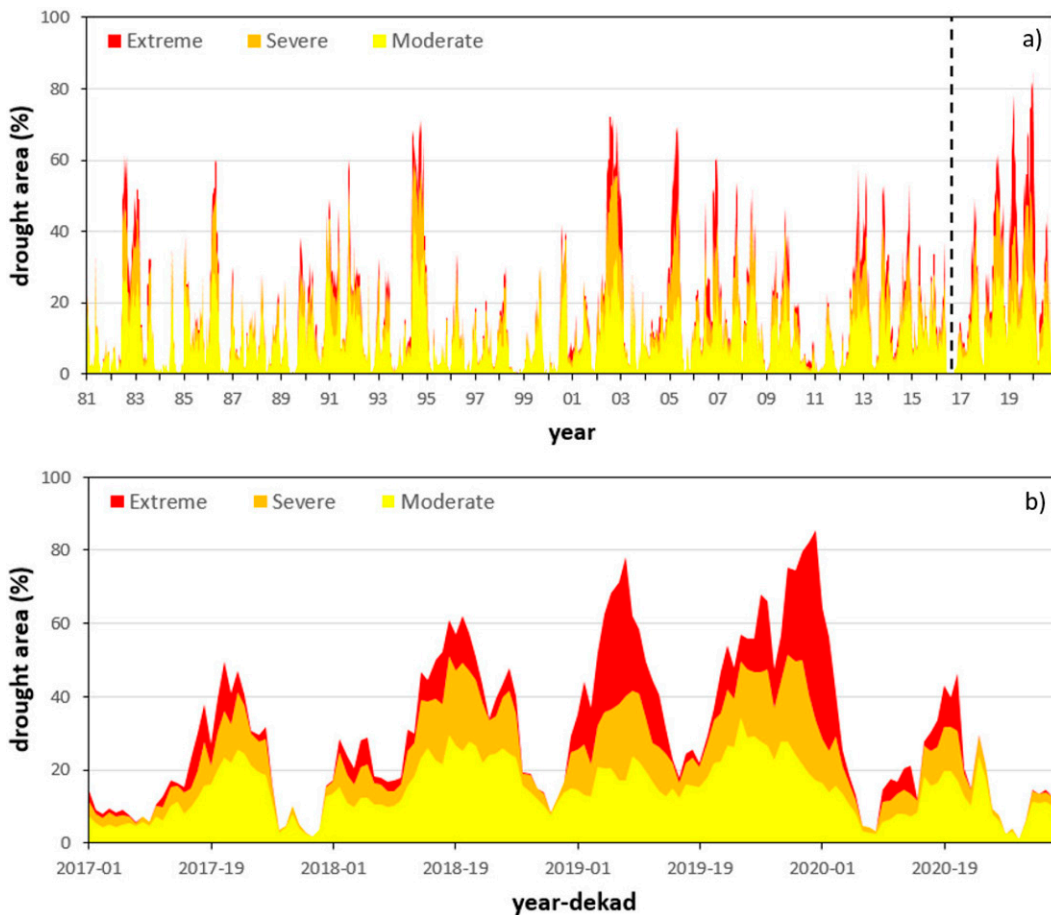


FIG. 7. Time series of the areas under drought conditions (in three severity categories, according to McKee et al. 1993) for mainland Australia according to SPI-3. (a) The full time series 1981–2020, with the dotted line demarking the start of the (b) study period in 2017, which is depicted in detail.

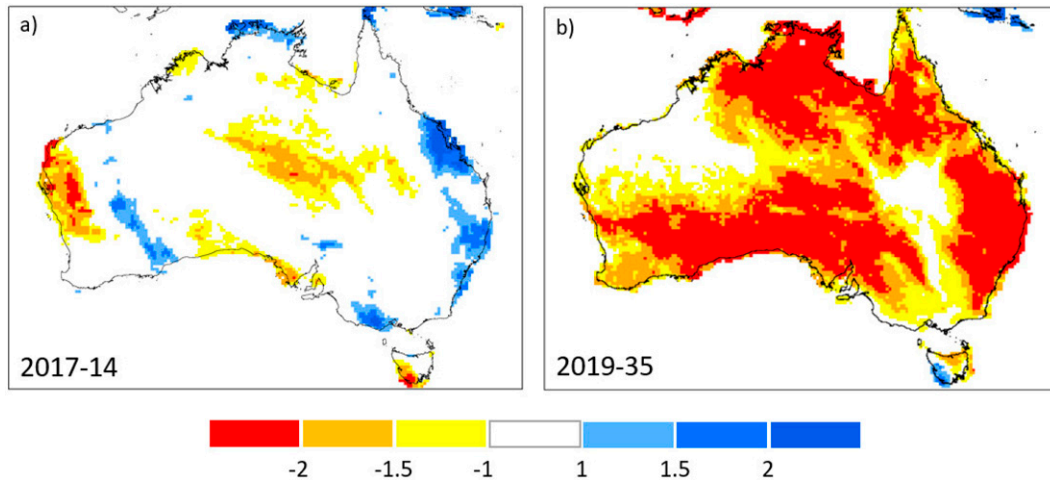


FIG. 8. SPI-3 maps at different phases of the study period. (a) The drought conditions in the middle of May 2017 (14th dekad of the year), at the transition between the rainy season in the north and in the south of the country. (b) The drought conditions at the end of 2019 (35th dekad of the year) spreading across the full country.

gridded data with dekad updates, and can be considered valid for global scale applications based on datasets with similar characteristics. Gridded data with higher/lower spatial resolution or temporal updates (i.e., monthly rather than dekadal) would require other optimized parameterizations that may lie outside the identified ranges (Table 1). The specific model parameterization depends, as already stressed in the model description, on the aim of the application, as further discussed in section 4b.

b. Case study

To evaluate the flexibility of the proposed methodology in adapting to potentially different characterizations of the same drought event, a case study is here investigated: the Australia drought period 2017–19. According to the Bureau of Meteorology of the Australian Government (<http://www.bom.gov.au/climate/drought/knowledge-centre/previous-droughts.shtml>) the years 2017–19 were unprecedented in terms of dry conditions during the cold season, with very little recovery during the months October–December between these consecutive dry periods.

The analysis of the full 1981–2020 time series of the areas under drought [in three severity categories, according to McKee et al. (1993)] according to SPI-3 clearly shows how this period was indeed exceptional (Fig. 7a), especially in terms of the extent of the drought at the peak (more than 80% of mainland Australia simultaneously under drought conditions). While the Australian Bureau of Meteorology considers this period a single drought event, with a combined rainfall deficit in the 3-yr period (January 2017–December 2019) being the lowest on record in many regions, differences in both severity and extent of the drought can be observed. Although all three April–September periods were among the driest on record for southeastern Australia, the 2019 cool season was drier than those of 2017 and 2018. These differences

are clearly shown in Fig. 7b, where 4–5 different drought periods can be detected.

In terms of the spatial distribution, the most extreme conditions were recorded in southeastern Australia, mainly the northern half of New South Wales and the eastern Victoria, while western Australia was mostly interested by drought conditions only earlier in the period. In 2017 there was a clear split between two droughts in the northeast and the southwest of the country, with dry conditions occurring mainly during the corresponding rainy seasons (October–April for the north and April–November for the south) with only a partial overlapping period (Fig. 8a). The droughts in 2018 and 2019 were more spread out and difficult to separate (e.g., Fig. 8b).

This overview of the 2017–19 Australian drought points to the need of adapting the event classification approach to the aim of the specific analysis an expert/use may run. In longer time-scale climate analysis, for instance, this event can be classified as a single drought event. However, shorter time-scale studies aiming at investigating differences within this drought period may split the full event into multiple shorter ones. To highlight the flexibility of the proposed algorithm to adapt to these different classifications of the same drought period, three different settings were used to analyze the 2017–19 Australian drought (Table 2).

The first parameterization of the model (ID = 1) addresses longer time-scale climate analyses; thus, it uses a small value for the minimum cluster size (A), the classic contiguous area with Boolean drought conditions in space [corresponding to

TABLE 2. Different parameterizations of the generalized model used to characterize the 2017–19 Australian drought period.

| ID | L | p | k | E | A | R |
|----|-----|------|------|------|-----|-----|
| 1 | 3 | 0.05 | −1 | 1000 | 65 | 5 |
| 2 | 3 | 0.05 | −1 | 1000 | 200 | 3 |
| 3 | 3 | 0.30 | −1.5 | 8 | 200 | 3 |

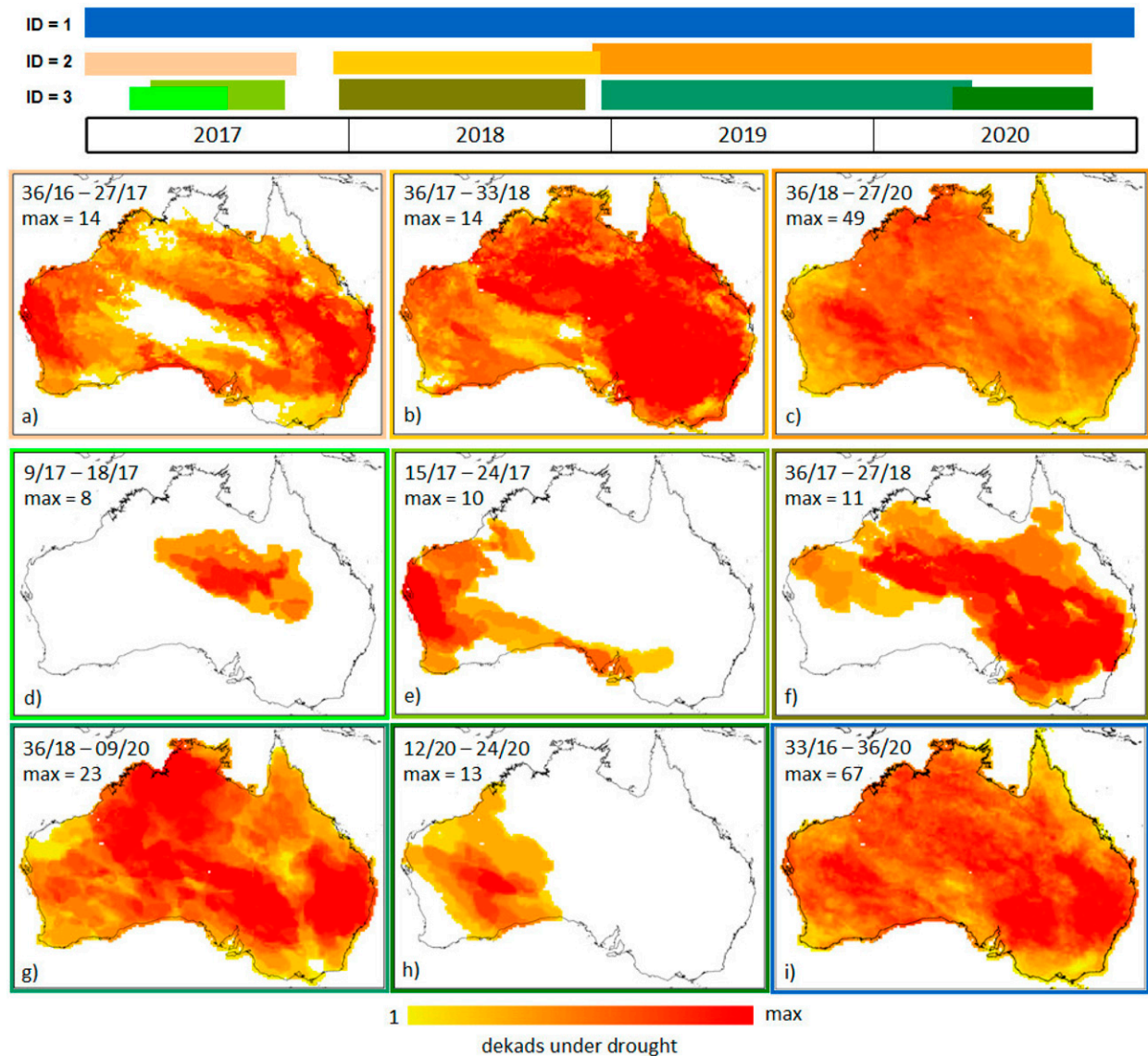


FIG. 9. Heat maps representing the number of time steps when each cell is included in a specific drought event according to the three different parameterizations (see Table 2). The bars at the top of the figure highlight the temporal extent of each event, with each duration also detailed by the start and end dekad and year (DD/YY) in the upper-left corner of each panel. (a)–(c) The results for the parameterization ID = 2 (boxed in orange shades), (d)–(h) the results for ID = 3 (boxed in green shades), and (i) the results for ID = 1 (boxed in blue).

the classical CDA proposed by [Andreadis et al. (2005)] and a larger search window in time than in space ($R > L$). The second set of parameters (ID = 2) increases the minimum cluster size (hence reducing the potential connections between large clusters), and adopts a symmetric space–time search window ($R = L$). Finally, the third parameterization (ID = 3) further increases the spatial separation by increasing the number of points needed to define a core point (i.e., increasing p) and by also replacing the simple Boolean weighting with the logistic function.

The outcome of the three clustering algorithms can be synthesized with heat maps, depicting how long each cell was under drought conditions during a given event (Fig. 9). The bars

at the top of Fig. 9 highlight how the full period is modeled as a single event, according to ID = 1, starting in December 2016 and ending in November 2020. The second set of parameters (ID = 2, orange shades) returns three main events: 2017, 2018, and 2019/20. These three events mostly correspond to the three main periods observable in Fig. 7, with the exception of the minor drought in 2020 that is modeled as part of the 2019 event. Finally, the third setting (ID = 3, green shades) identifies five main events: 2018, 2019, 2020, and two partially overlapping events in 2017. As complementary information, the plots in Fig. 10 track the spatiotemporal evolution of the centroids of the events according to the different parameterization.

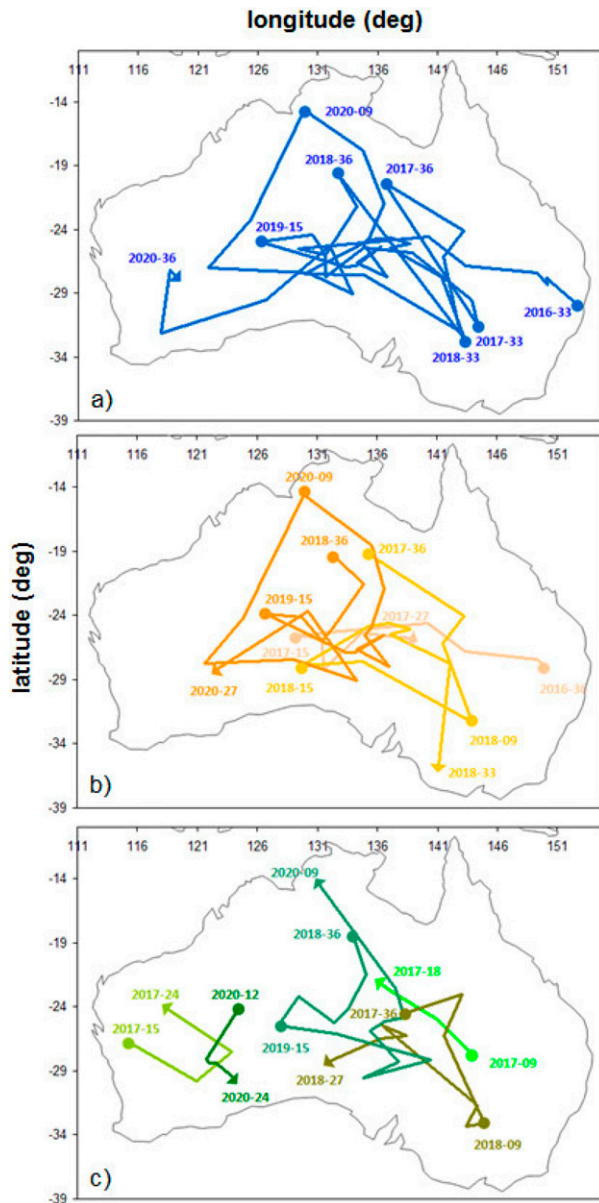


FIG. 10. Spatiotemporal tracking of the centroids of the events according to the three different parameterizations of the generalized clustering algorithm (see Table 2). The colors of the tracks match the ones used in Fig. 9. Start and end dates, as well as some key dates, are reported to better identify the temporal evolution.

These two sets of information provide a rather different narrative for the full drought period, depending on the selected parameterization. In the $ID = 2$ case, the 2017 and 2018 events (Figs. 9a and 9b, respectively) have similar duration, even if the second one has a much larger spatial spread, while the former has a more marked east to west movement (Fig. 10b). The event in 2019/20 (Fig. 9c) is much longer than the other two, and with a less marked difference between eastern and western Australia. Due to its extent, the centroid of the event stays around central Australia most of the time.

Differently, the $ID = 3$ case clearly shows the absence of spatial overlap between the two events in northeastern (Fig. 9d) and southwestern (Fig. 9e) Australia in 2017, which is clearly noticeable also in the spatial separation of the paths of the centroids in Fig. 10c. It also highlights the strongest drought conditions in 2019 (Fig. 9g) over New South Wales but also the Northern Territory (which were masked by the inclusion of 2020 in the previous case). According to the $ID = 3$ parameterization, this event starts and ends in the Northern Territory (Fig. 10c), while moving to southeast during the peak phase.

The $ID = 1$ case provides a synthetic summary of the full drought period, showing how the New South Wales was the region most affected by the drought, mainly due to the 2018 and 2019 drought periods (Fig. 9i), but it does not allow to make further detailed considerations. However, the $ID = 1$ parameterization is very useful to contextualize the full period in the historical framework of other long drought events in the area, such as the Millennium drought, as well as to highlight the overall movement of the drought along the full period (Fig. 10a). In this regard, it is possible to observe a general tendency of this event to move from east to west, even if north-to-south migration can be observed in two occasions at the end of 2017 and 2018, corresponding to the period of minimum extent of the area under drought (see Fig. 7).

Finally, Fig. 11 shows the temporal evolution of the area under drought conditions according to SPI-3 (Figs. 11a,c,e) and SPI-6 (Figs. 11b,d,f), obtained by applying the same three sets of parameters reported in Table 2. For both indices, the outcome of the three settings is in line with the expected behavior described in the previous sections, with a stronger connectivity between clusters (i.e., fewer and larger events) when low A and large R values are used (Figs. 11a,b). On the opposite, more (and smaller) clusters are obtained when the logistic weighting function is used together with higher p and A values (Figs. 11e,f).

Even if the overall behavior of the algorithm seems to be coherent for the two SPI datasets, some discrepancies in the results can be observed, especially for the parameterization $ID = 2$ where only two events are modeled for SPI-6 compared to the three events reported for SPI-3. These differences can be ascribed to the smoother dynamics of SPI-6 compared to SPI-3 due to the longer accumulation period. They also suggest that finetuning of the proposed optimal parameterization ranges may be needed for specific indicators, even if the rules introduced in the previous sections remain of general validity.

5. Summary and conclusions

In this study, we introduced a generalized algorithm for the spatiotemporal tracking of drought events at a global scale. Starting from the contiguous drought area methodology (Andreadis et al. 2005), we contextualized the proposed approach within the data-mining framework of a three-dimensional implementation of the density-based DBSCAN clustering method. This implementation represents a generalization of the CDA approach which is a special case of the proposed

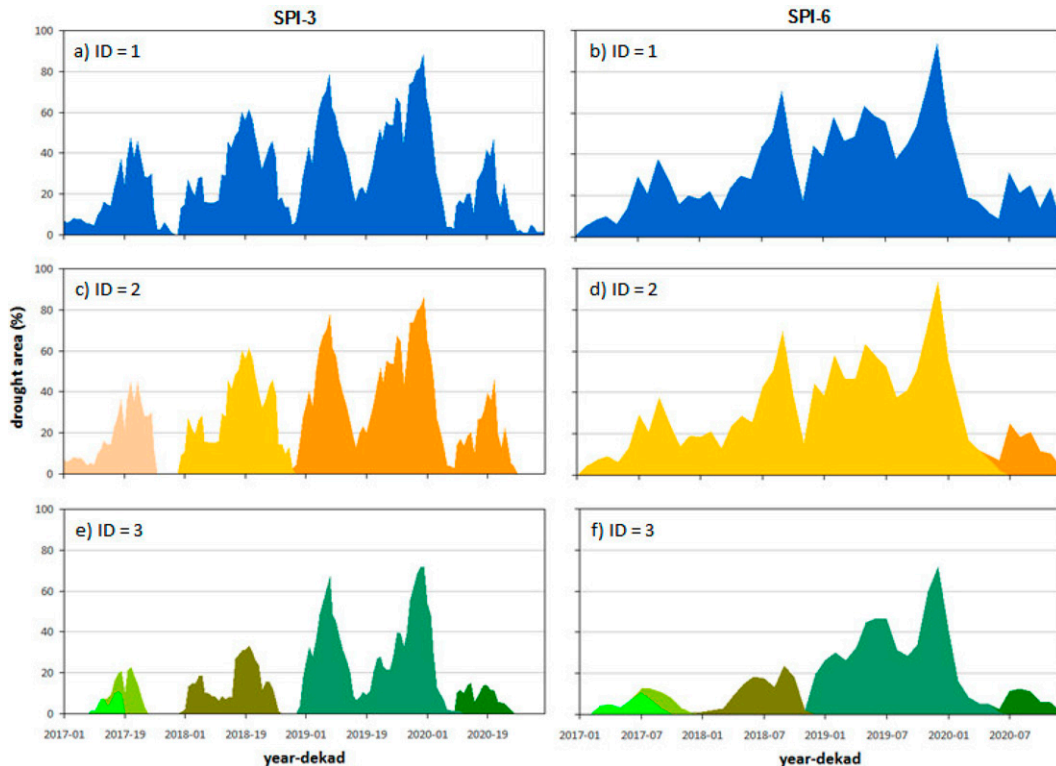


FIG. 11. Temporal evolution of the 2017–19 Australian drought according to the three different parameterizations of the generalized clustering algorithm (see Table 2). (a),(c),(e) The time series of drought areas based on SPI-3 data. (b),(d),(f) The drought areas based on SPI-6. Different shades represent different clusters according to the method.

method. It is based on a set of six parameters that allows for an increased flexibility of the algorithm in both spatial and temporal connectivity of cells under drought, which expands the possibility of adapting to different definitions of drought events.

The optimal range of variability of each parameter was investigated on a medium-resolution SPI-3 dataset (derived from the ECMWF reanalysis ERA5 precipitation), by showing the effects on the clustering in some exemplary cases. Although this represents a rule-of-thumb approach, to guide potential users in selecting the optimal parameterization according to the specific needs of their applications, it provides key information on how the parameters affect the onset, propagation, and end of a drought event.

The proposed approach is designed as a 3D clustering, hence it requires a full data cube (covering both the spatial and temporal domains) for its application. This is not a constraint for historical analyses, but it may provide an additional challenge in the case of near-real-time applications. In this case, newly acquired data may alter the outcome of the clustering output, propagating even far in the past. The application of the proposed methodology with a suitable temporal lag (e.g., 1–2 temporal steps, depending on the R parameter) should limit these negative effects and provide stable outcomes, but these assumptions need to be properly tested.

The actual capability of the generalized algorithm to adapt to different characterizations of the same drought event was

tested for a case study in Australia using SPI-3 data updated every 10 days. Three different parameterizations of the algorithm returned rather different descriptions of the spatiotemporal evolution of the event, ranging from a single multiannual event to multiple seasonal ones. This general behavior remains consistent even if an index with a different accumulation period and frequency is used (monthly SPI-6), even if different clusters are returned. This result reveals how the algorithm can be tuned to fulfill the needs and goals of various users. Even if rather different outcomes can be obtained for the same event, once an indicator and a parameterization are selected (based on the specific objectives a study targets) the algorithm can provide consistent estimations of different drought events at global scale, allowing for objective comparisons among events without any additional a priori knowledge.

Data availability statement. All the data used in this study can be accessed at the Copernicus Global Drought Observatory (GDO) web portal (<https://edo.jrc.ec.europa.eu/gdo/php/index.php?id=2112>). The source code of the clustering algorithm, developed in the open-source language R, can be also requested via GDO.

REFERENCES

Alley, W. M., 1984: The palmer drought severity index: Limitations and assumptions. *J. Appl. Meteor. Climatol.*, **23**, 1100–1109,

- [https://doi.org/10.1175/1520-0450\(1984\)023<1100:TPDSIL>2.0.CO;2](https://doi.org/10.1175/1520-0450(1984)023<1100:TPDSIL>2.0.CO;2).
- Andreadis, K. M., E. A. Clark, A. W. Wood, A. F. Hamlet, and D. P. Lettenmaier, 2005: Twentieth-century drought in the conterminous United States. *J. Hydrometeorol.*, **6**, 985–1001, <https://doi.org/10.1175/JHM450.1>.
- Cammalleri, C., F. Micale, and J. V. Vogt, 2016: A novel soil moisture-based drought severity index (DSI) combining water deficit magnitude and frequency. *Hydrol. Processes*, **30**, 289–301, <https://doi.org/10.1002/hyp.10578>.
- , J. Spinoni, P. Barbosa, A. Toreti, and J. V. Vogt, 2022: The effects of non-stationarity on SPI for operational drought monitoring in Europe. *Int. J. Climatol.*, **42**, 3418–3430, <https://doi.org/10.1002/joc.7424>.
- Diaz, V., G. A. Corzo Perez, H. A. J. Van Lanen, D. Solomatine, and E. A. Varouchakis, 2020: An approach to characterise spatio-temporal drought dynamics. *Adv. Water Resour.*, **137**, 103512, <https://doi.org/10.1016/j.advwatres.2020.103512>.
- , —, —, and —, 2023: Three-dimensional clustering in the characterization of spatiotemporal drought dynamics: Cluster size filter and drought indicator threshold optimization. *Advances in Hydroinformatics: Artificial Intelligence and Optimization for Water Resources*, D. Solomatine and G. A. Corzo Perez, Eds., Wiley, in press.
- Ester, M., H.-P. Kriegel, J. Sander, and X. Xu, 1996: A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD'96: Proc. Second Int. Conf. on Knowledge Discovery and Data Mining*, Portland, OR, AAAI Press, 226–231, <https://dl.acm.org/doi/10.5555/3001460.3001507>.
- González-Hidalgo, J. C., S. M. Vicente-Serrano, D. Peña-Angulo, C. Salinas, M. Tomas-Burguera, and S. Beguería, 2018: High-resolution spatio-temporal analyses of drought episodes in the western Mediterranean basin (Spanish mainland, Iberian Peninsula). *Acta Geophys.*, **66**, 381–392, <https://doi.org/10.1007/s11600-018-0138-x>.
- Hahsler, M., M. Piekenbrock, and D. Doran, 2019: DBSCAN: Fast density-based clustering with R. *J. Stat. Software*, **91** (1), 1–30, <https://doi.org/10.18637/jss.v091.i01>.
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Herrera-Estrada, J. E., Y. Satoh, and J. Sheffield, 2017: Spatio-temporal dynamics of global drought. *Geophys. Res. Lett.*, **44**, 2254–2263, <https://doi.org/10.1002/2016GL071768>.
- Kriegel, H.-P., P. Kröger, J. Sander, and A. Zimek, 2011: Density-based clustering. *WIREs Data Mining Knowl. Discovery*, **1**, 231–240, <https://doi.org/10.1002/widm.30>.
- Lloyd-Hughes, B., 2012: A spatio-temporal structure-based approach to drought characterization. *Int. J. Climatol.*, **32**, 406–418, <https://doi.org/10.1002/joc.2280>.
- Madhulatha, T. S., 2012: An overview of clustering methods. *IOSR J. Eng.*, **2**, 719–725, <https://doi.org/10.9790/3021-0204719725>.
- McKee, T. B., N. J. Doesken, and J. Kleist, 1993: The relationship of drought frequency and duration to time scales. *Eighth Conf. on Applied Climatology*, Anaheim, CA, Amer. Meteor. Soc., 179–184.
- Mishra, A. K., and V. P. Singh, 2010: A review of drought concepts. *J. Hydrol.*, **391**, 202–216, <https://doi.org/10.1016/j.jhydrol.2010.07.012>.
- Rakovec, O., L. Samaniego, V. Hari, Y. Markonis, V. Moravec, S. Thober, M. Hanel, and R. Kumar, 2022: The 2018–2020 multi-year drought sets a new benchmark in Europe. *Earth's Future*, **10**, e2021EF002394, <https://doi.org/10.1029/2021EF002394>.
- Sheffield, J., K. M. Andreadis, E. F. Wood, and D. P. Lettenmaier, 2009: Global and continental drought in the second half of the twentieth century: Severity-area-duration analysis and temporal variability of large-scale events. *J. Climate*, **22**, 1962–1981, <https://doi.org/10.1175/2008JCL12722.1>.
- Spinoni, J., G. Naumann, J. V. Vogt, and P. Barbosa, 2015: The biggest drought events in Europe from 1950 to 2012. *J. Hydrol. Reg. Stud.*, **3**, 509–524, <https://doi.org/10.1016/j.ejrh.2015.01.001>.
- Stasinopoulos, D. M., and R. A. Rigby, 2007: Generalized additive models for location scale and shape (GAMLSS). *J. Stat. Software*, **23** (7), 1–46, <https://doi.org/10.18637/jss.v023.i07>.
- Tallaksen, L. M., and H. A. J. Van Lanen, 2004: Drought as a natural hazard: Introduction. *Hydrological Drought Processes and Estimation Methods for Streamflow and Groundwater, Developments in Water Science*, L. M. Tallaksen and H. A. M. Van Lanen, Eds., Elsevier, 3–17.
- Tilloy, A., B. D. Malamud, and A. Joly-Laugel, 2022: A methodology for the spatiotemporal identification of compound hazards: Wind and precipitation extremes in Great Britain (1979–2019). *Earth Syst. Dyn.*, **13**, 993–1020, <https://doi.org/10.5194/esd-13-993-2022>.
- UNDRR, 2021: GAR special report on drought 2021. United Nations Office for Disaster Risk Reduction, 210 pp., <https://www.undrr.org/publication/gar-special-report-drought-2021>.
- Vernieuwe, H., B. De Baets, and N. E. C. Verhoest, 2020: A mathematical morphology approach for a qualitative exploration of drought events in space and time. *Int. J. Climatol.*, **40**, 530–543, <https://doi.org/10.1002/joc.6226>.
- Wang, J., and Z. Yan, 2021: Rapid rises in the magnitude and risk of extreme regional heat wave events in China. *Wea. Climate Extremes*, **34**, 100379, <https://doi.org/10.1016/j.wace.2021.100379>.
- WMO, 2017: Guidelines on the calculation of climate normals. WMO/TD-1203, 29 pp., https://library.wmo.int/doc_num.php?explnum_id=4166.
- , and GWP, 2016: Handbook of drought indicators and indices. WMO/TD-1173, 45 pp., https://library.wmo.int/doc_num.php?explnum_id=3057.
- Zhao, S., D. Cong, K. He, H. Yang, and Z. Quin, 2017: Spatial-temporal variation of drought in China from 1982 to 2010 based on a modified temperature vegetation drought index (mTVDI). *Sci. Rep.*, **7**, 17473, <https://doi.org/10.1038/s41598-017-17810-3>.